INDIAN INSTITUTE OF TECHNOLOGY ROORKEE

Certificate course on AI and ML

Exploratory Data Analysis

Prof. Kusum Deep
Full Professor (HAG), Department of Mathematics
Joint Faculty, MF School of Data Science and Artificial Intelligence
Indian Institute of Technology Roorkee, Roorkee – 247667

kusum.deep@ma.iitr.ac.in, kusumdeep@gmail.com

Learning outcomes

- Definition of data and understand its importance
- Data Analytics and its various types
- Why Analytics is important in today's world
- Relationship between statistics, analytics and data science
- Levels of Data

Exploratory Data Analysis

Def: Exploratory Data Analysis (EDA) is an approach to **analyzing** and **understanding** data sets in order to **summarize** their main characteristics, often with the help of **visualization** and **summary statistics**. The main goal of EDA is to uncover **patterns** and **relationships** that may be hidden within the data, and to provide **insights** into the underlying structure of the data.

It involves following techniques:

- data visualization,
- summary statistics, and
- hypothesis testing.

It can be used to identify **outliers**, **missing values**, **trends**, and other **patterns** in the data, and to help guide the selection of **appropriate statistical models** and methods for further analysis. It is used in the early stages of a data analysis project, to gain a deeper understanding of the data and to generate new ideas for analysis.

Definition of Data and its importance

Data is a collection of information. It is the raw material used to generate insights, understand patterns, and inform decision-making.

Examples: numbers, text, images.

Without data, there can be no analysis.

The quality, quantity, and accuracy of the data can significantly impact the results of the analysis.

Def: **Structured Data**

Data that is organized in a specific, predefined format, making it easier to analyze and process by computer programs. Example are:

Relational database tables - data stored in tables with columns and rows that have a specific data type and structure, such as SQL databases.

Spreadsheets - data stored in rows and columns in software like Microsoft Excel or Google Sheets.

Sensor data - data collected from IoT devices or other sensors that have a specific structure, such as temperature, humidity, or pressure readings.

Financial data - data related to financial transactions, such as bank statements, credit card transactions, or stock market data.

E-commerce data - data related to online purchases, such as order details, customer information, and product information.

Medical records - data related to patient information, such as diagnosis, medication history, and medical test results.

Structured data is easier to analyze and process compared to unstructured data since it has a predefined format and structure. This is why structured data is commonly used in data analysis and machine learning applications.

Def: **Unstructured Data**

It does not have a specific, predefined format, making it difficult to analyze and process by computer programs. Examples:

Text data - data in the form of natural language text, such as emails, social media posts, news articles, or blog posts.

Images and videos - data in the form of visual media, such as photographs, videos, or animations.

Audio data - data in the form of sound recordings, such as music, podcasts, or phone call recordings.

Webpage content - data from web pages, such as HTML code, metadata, or web logs.

Sensor data - data from sensors that produce unstructured data, such as audio sensors or video sensors.

Social media data - data from social media platforms, such as tweets, comments, or shares.

Unstructured data is more challenging to analyze and process than structured data since it lacks a specific format and structure. However, it contains valuable insights that can be extracted using natural language processing (NLP), image recognition, or other techniques. With the help of machine learning, unstructured data can be transformed into structured data for further analysis.

What is generating so much data ?

Growth of digital technologies: The increasing use of digital technologies such as smartphones, tablets, and IoT has led to the generation of vast amounts of data. These devices generate data continuously, through sensors, applications, and other means.

Social media and online platforms: e.g. from user interactions, such aslikes, comments, and shares.

Business operations: Companies generate vast amounts of data through their day-to-day operations, such as sales data, customer information, and supply chain data.

Scientific research: generates large amounts of data, such as genomic data, environmental data, and astronomical data.

Digital transformation: The digitization of industries such as healthcare, finance, and manufacturing has led to the generation of large amounts of data. For example, electronic medical records generate vast amounts of patient data that can be analyzed for insights.

The growth in data generation is expected to continue as technology continues to advance, leading to the need for advanced tools and techniques to store, manage, and analyze data effectively.

How data add value to business?

Better decision-making: Data-driven decision-making allows businesses to make informed decisions based on accurate and relevant data, rather than relying on intuition or guesswork. This can lead to better outcomes and higher profitability.

Improved customer experience: Analyzing customer data can provide insights into customer behavior, preferences, and needs, allowing businesses to improve customer experience and satisfaction.

Operational efficiency: Data analysis can help identify inefficiencies in business operations, allowing businesses to optimize processes and improve productivity.

Competitive advantage: By leveraging data to gain insights into market trends, customer behavior, and competitor activities, businesses can gain a competitive advantage and differentiate themselves in the marketplace.

Innovation: Data can be used to identify new opportunities for innovation and product development, leading to the creation of new products and services that meet customer needs.

Risk management: Analyzing data can help businesses identify potential risks and vulnerabilities, allowing them to take steps to mitigate these risks and protect the business.

Data products

Data products are information-based products that are created using data and analytics. They can be physical or digital and are designed to help individuals or organizations make better decisions, improve processes, or create new products or services. Examples :

Personalized recommendations: They use customer data to provide personalized product recommendations for online shoppers, such as Netflix's recommendation system.

Real-time analytics dashboards: They provide real-time insights into business metrics, such as sales revenue or website traffic, using interactive visualizations.

Predictive models: They use machine learning algorithms to predict future outcomes based on historical data, such as predicting customer churn or forecasting sales revenue.

Chatbots and virtual assistants: They use natural language processing to understand and respond to customer inquiries, such as booking a flight or ordering food.

Fraud detection systems: They use machine learning algorithms to identify patterns of fraudulent activity, such as credit card fraud or insurance claims fraud.

Smart devices: They use sensors and internet connectivity to collect and analyze data, such as smart thermostats, fitness trackers, or home security systems.

Geospatial analytics: They use geographic data to analyze trends and patterns, such as analyzing traffic patterns or predicting weather patterns.

Why data is important?

Data is crucial in enabling organizations and individuals to make informed decisions, improve efficiency, enhance customer experiences, manage risks, identify underperformance, understand customers, understand market and drive innovation.

Define data analytic and its types

- Define data analytics
- Why analytics is important?
- Data analysis
- Data analytics vs. Data analysis
- Types of Data analytics

Define data analytics

Data analytics is the process of collecting, cleaning, processing, and analyzing large sets of data to identify patterns, trends, and insights. It involves using various statistical and computational techniques to extract meaningful information from data and make informed decisions based on that information.

It is performed using the following steps:

Data collection

Data cleaning

Data processing

Data analysis

Data visualization

It can be applied to various fields such as business, healthcare, finance, and science. It can be used to inform decision-making, improve processes, and gain a competitive advantage.

Data analysis

Data analysis is the process of examining and interpreting data in order to extract meaningful insights, identify patterns and trends, and make informed decisions. It is used in business, healthcare, science, and government. It is done in steps:

- **Data collection**
- **Data cleaning**
- **Data processing**
- **Data exploration**
- **Data visualization**
- **Drawing conclusions and making recommendations**

It can be performed using Excel spreadsheets, R or Python.

Is Analysis = Analytics ?

Analysis is a subset of **analytics**. Analytics involves the entire process of collecting, managing, analyzing, and visualizing data, while analysis is a specific part of that process that focuses on examining data to identify patterns and relationships.

Why analytics is important?

Analytics is important because it provides organizations with the information they need to make data-driven decisions, improve operations, and stay competitive in an ever-changing business environment. Fo example:

- Determining credit risk
- Developing new medicines
- Finding more efficient ways to deliver products and services
- Preventing fraud
- Uncovering cyber threats
- Retaining the most valuable customers

Example

Consider a retail company

Data analysis may involve examining sales data to identify patterns and trends, such as which products are selling well and which are not, and what factors are contributing to those sales trends. This might involve using statistical analysis to identify correlations between different data sets, such as sales data and customer demographics, to help the company make informed decisions about pricing, inventory, and marketing.

Data analytics would involve a more comprehensive approach to data management and analysis. In addition to analyzing sales data, data analytics might involve collecting data from multiple sources, such as social media, website traffic, and customer feedback, to gain a more complete understanding of customer behavior and preferences. This might involve using advanced tools such as machine learning algorithms and big data platforms to process and analyze large and complex data sets, and using data visualization techniques to communicate insights to stakeholders.

Classification of Data analytics

There are four classifications

1. **Descriptive analytics:**

It is the most basic form of analytics, which involves examining historical data to identify patterns, trends, and insights. This type of analytics is useful for understanding past performance and making informed decisions based on that information.

Examples of descriptive analytics include basic statistical analysis and data visualization.

2. Diagnostic analytics:

It involves analyzing data to identify the root causes of problems or issues. This type of analytics is useful for understanding why certain events occurred and how to prevent them from happening in the future.

Examples of diagnostic analytics include regression analysis and root cause analysis.

3. Predictive analytics:

It involves using statistical and machine learning algorithms to analyze data and make predictions about future events. This type of analytics is useful for identifying trends and patterns that can help inform future decision-making.

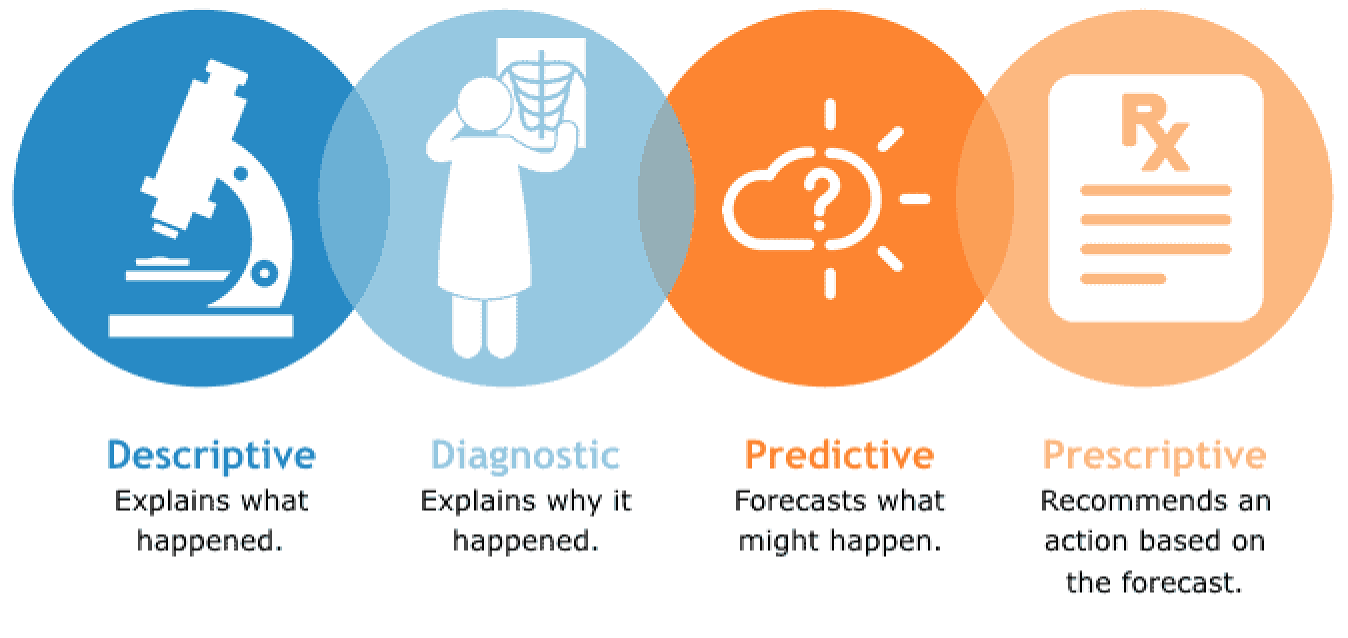
Examples of predictive analytics include forecasting and predictive modeling.

4. Prescriptive analytics:

It involves using data and algorithms to optimize decision-making and outcomes. This type of analytics is useful for identifying the best course of action in a given situation, based on various factors and constraints.

Examples of prescriptive analytics include optimization modeling and decision analysis.

Thus, the different types of data analytics build upon one another, starting with descriptive analytics to understand past performance, moving on to diagnostic analytics to identify the root causes of problems, and then using predictive and prescriptive analytics to inform future decision-making and optimize outcomes.



Descriptive
Explains what happened.

Diagnostic
Explains why it happened.

Predictive
Forecasts what might happen.

Prescriptive
Recommends an action based on the forecast.

Exs of Descriptive Analytics

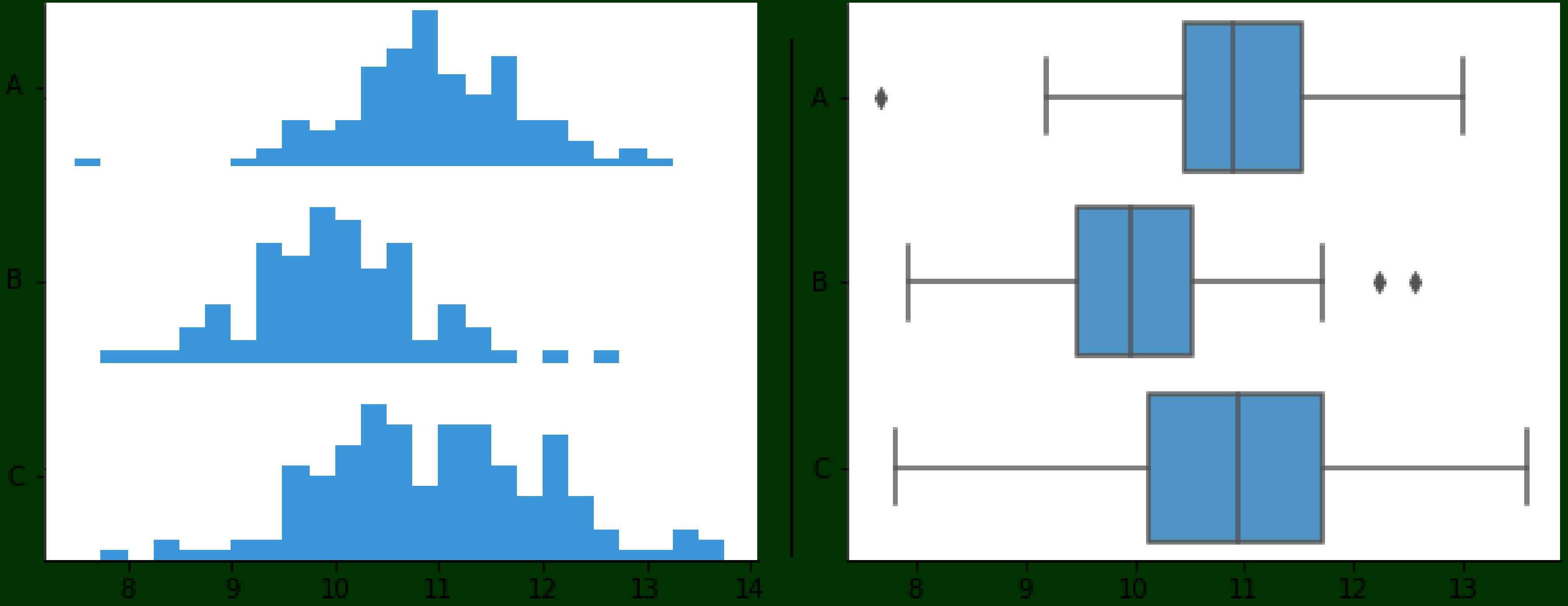
It is conventional form which provides a “summary view” which can describe event(s) that has occurred in the past.

Ex:

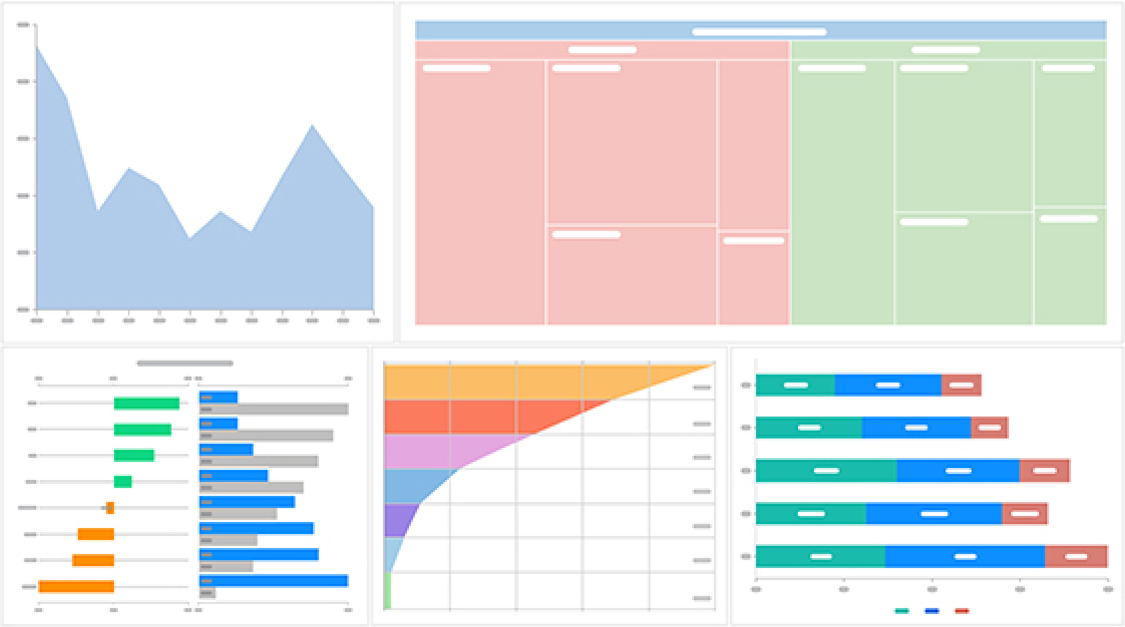
- Suppose we want to analyze the distribution of scores and understand the average performance of the students from a given dataset of students' test scores.

Create a histogram and box plot to identify outliers.

Ex 1



Ex 2



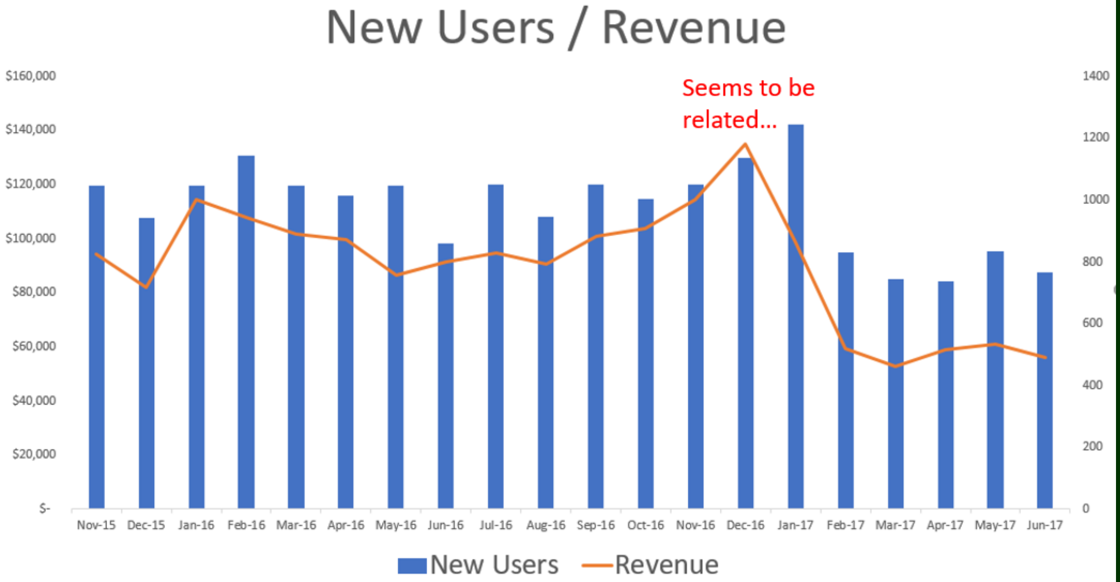
Top 5 Business Graphs and Charts Examples

Diagnostic analytics

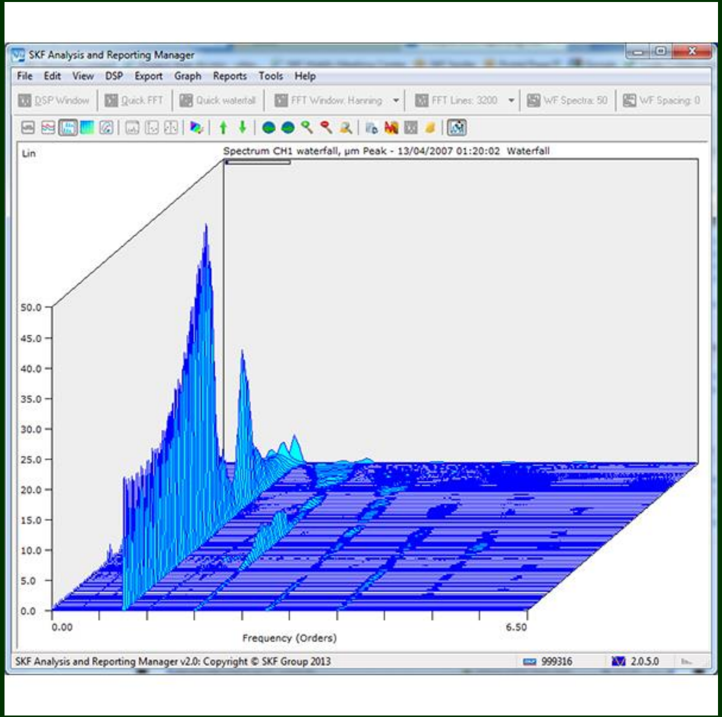
It is advanced analytics which answers “Why did it happen?”

May use

1. Data Discovery
2. Data Mining
3. Correlations



Ex 3



SKF Analysis and Reporting Manager

FFT Window: Hanning

FFT Lines: 3200

WF Spectra: 50

WF Spacing: 0

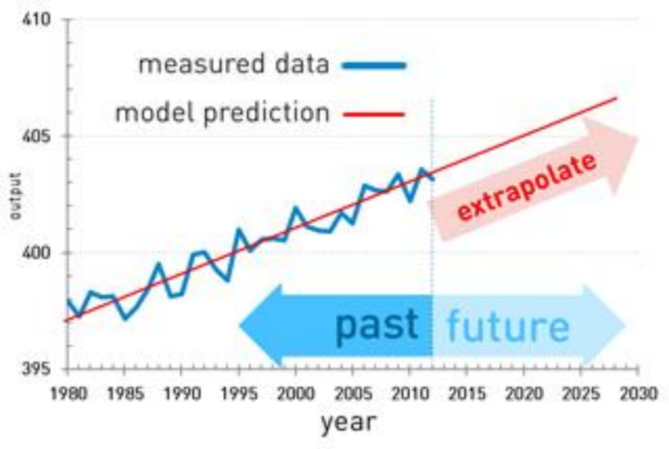
Spectrum CH1 waterfall, µm Peak - 13/04/2007 01:20:02 Waterfall

SKF Analysis and Reporting Manager v2.0: Copyright © SKF Group 2013

Predictive analytics

It helps to forecast trends, probability of an event happening in future or estimating the accurate time it will happen. Tools:

1. Linear regression
2. Time series analysis and forecasting
3. Data mining

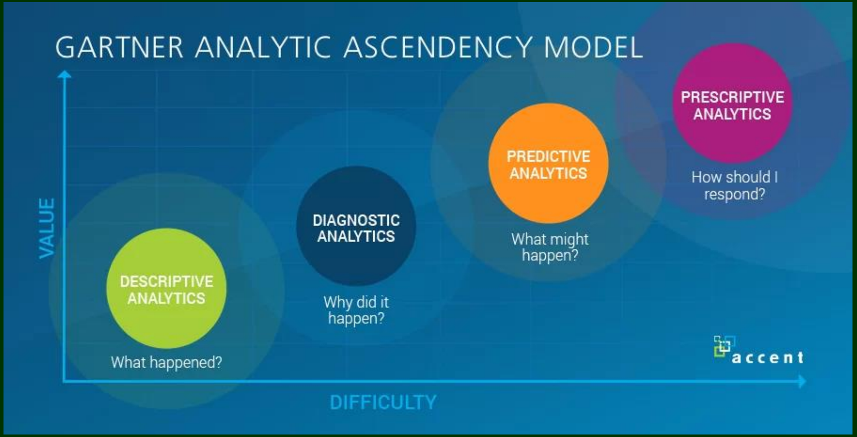


Ex 4

Prescriptive analytics

It is a set of techniques used to determine the best course of action for optimizing outcomes with an aim to improve quality, enhance services, reduce costs, and increase productivity.

- Optimization Model
- Simulation
- Decision Analysis



Ex 5

Why analytics is important

Analytics is the process of analyzing data using various statistical and computational methods to gain insights and make informed decisions.

It is important due to:

Data-driven decision making: Analytics enables organizations to make data-driven decisions, which can help them to identify new opportunities, reduce risks, and improve overall performance.

Performance measurement: Analytics allows businesses to measure and track their performance over time, which can help them to identify areas where they need to improve and make necessary changes.

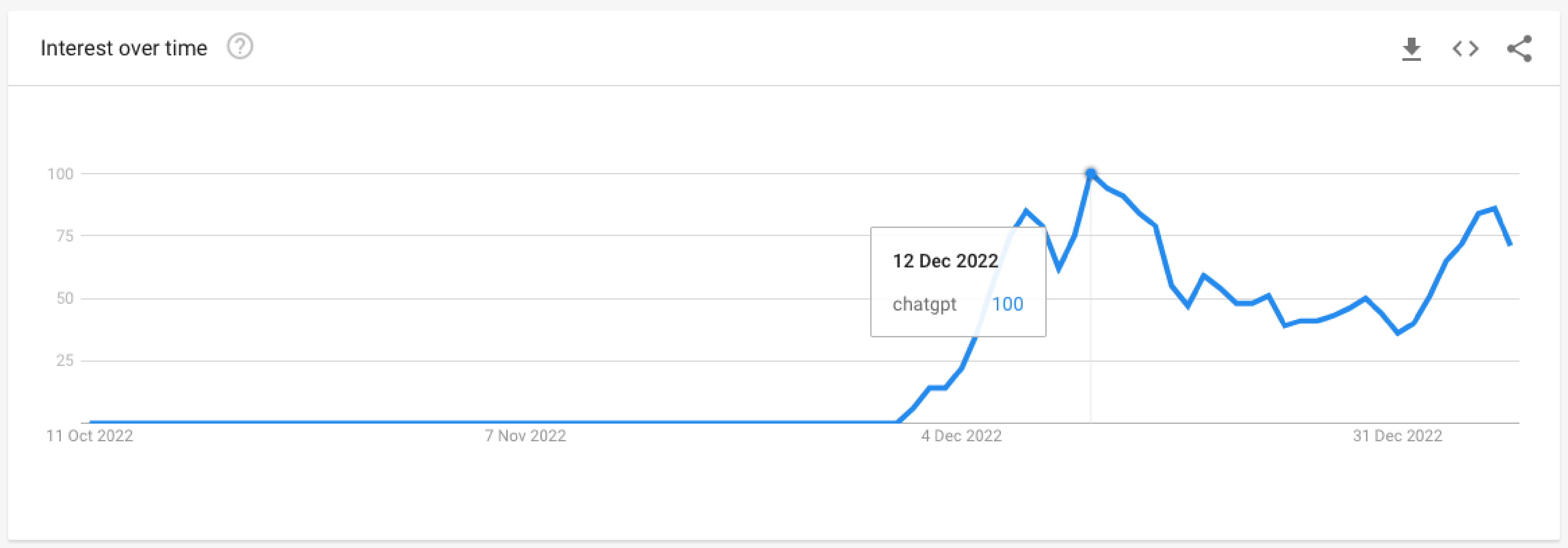
Competitive advantage: Analytics can provide businesses with a competitive advantage by helping them to identify trends and patterns in their data that their competitors may not have noticed.

Personalization: Analytics can be used to analyze customer behavior and preferences, which can help businesses to personalize their marketing campaigns and provide better customer experiences.

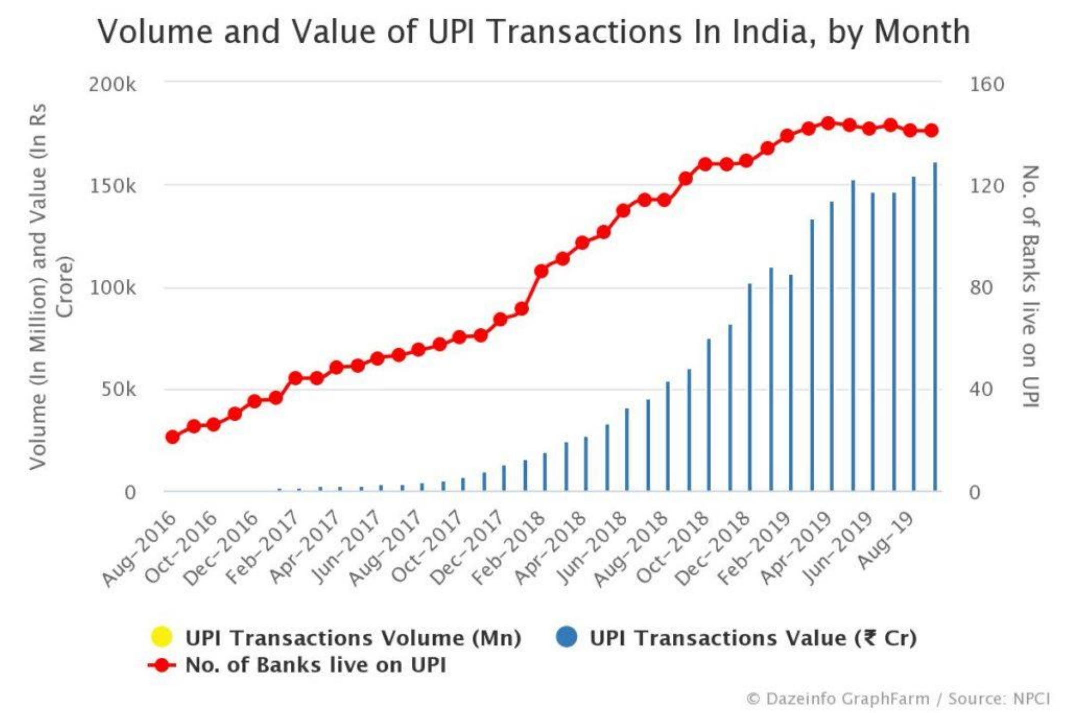
Cost reduction: Analytics can help businesses to identify areas where they can reduce costs, such as by optimizing their supply chain or improving their production processes.

Thus, analytics is important because it enables businesses to make better decisions, improve their performance, and gain a competitive advantage in their industry.

Ex: No. of ChatGPT users

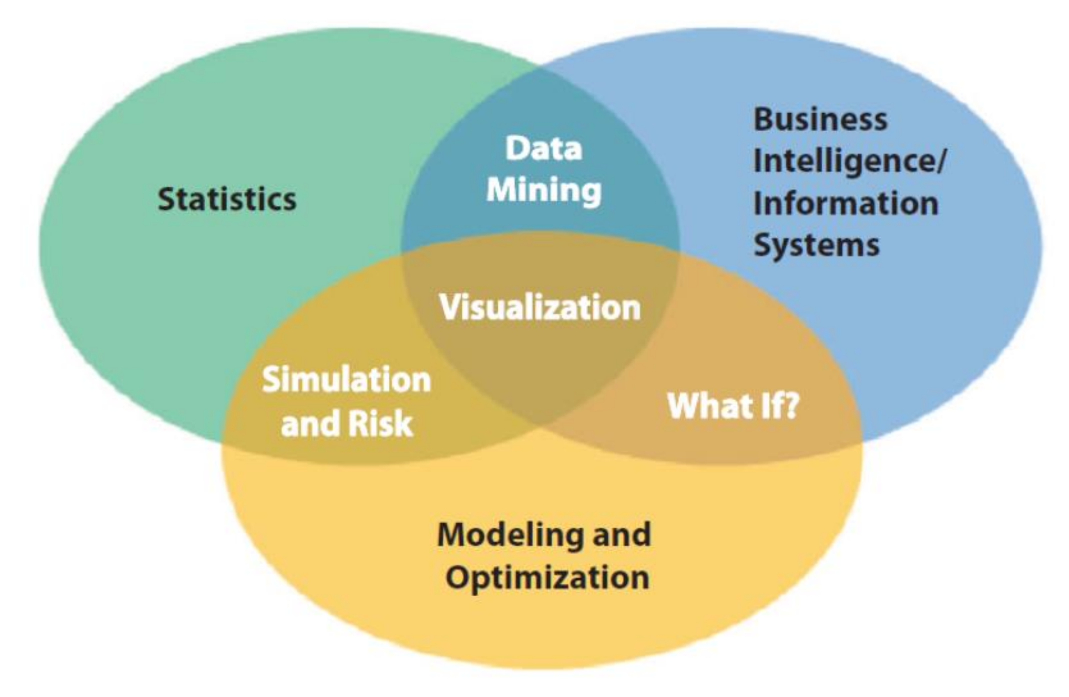


Ex: Volume and Value of UPI Transactions in India



© Dazeinfo GraphFarm / Source: NPCI

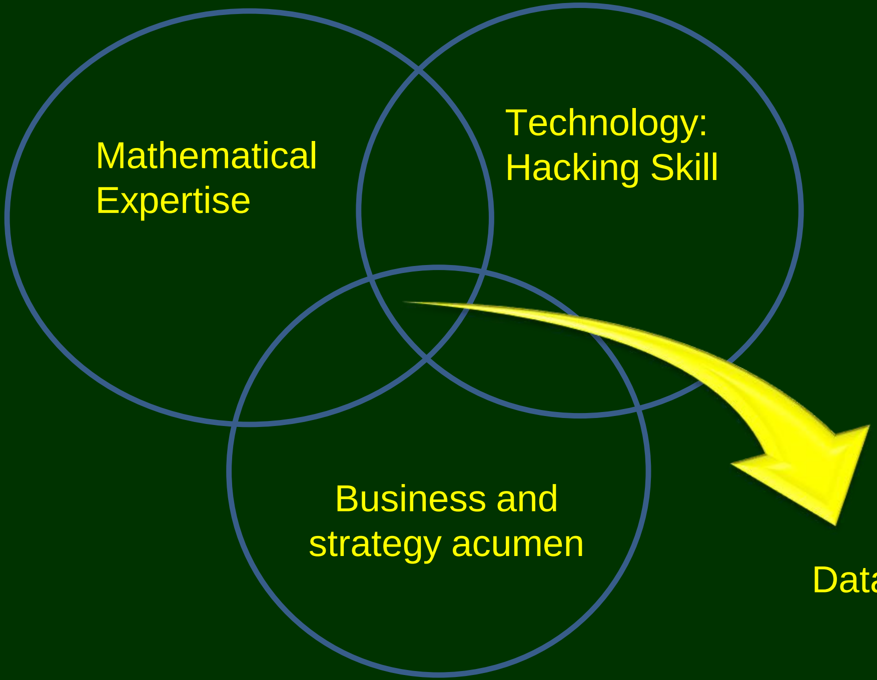
Element of data Analytics


Statistics
Data Mining
Business Intelligence/ Information Systems
Visualization
Simulation and Risk
What If?
Modeling and Optimization

Data analyst and Data scientist

- The requisite skill set
- Difference between Data analyst and Data Scientist

Required skill set



Mathematical Expertise

Technology: Hacking Skill

Business and strategy acumen

Data Science

Difference between Data analyst and Data Scientist

Job Scope: Data analysts are focused on analyzing data and generating insights that can be used to make business decisions, working with structured data and use tools like SQL, Excel, and Tableau to perform analysis.

Data scientists are involved in more complex data modeling tasks and developing machine learning algorithms, working with a wide variety of data types including structured, semi-structured and unstructured data.

Tools and Technologies: Data analysts use tools like SQL, Excel, and Tableau to perform analysis, while data scientists use a wider range of tools including programming languages such as Python or R, and machine learning frameworks such as TensorFlow, Keras or PyTorch.

Statistical Rigor: Data analysts use statistical techniques such as regression analysis and hypothesis testing, while data scientists use advanced statistical techniques such as machine learning algorithms, deep learning and artificial intelligence.

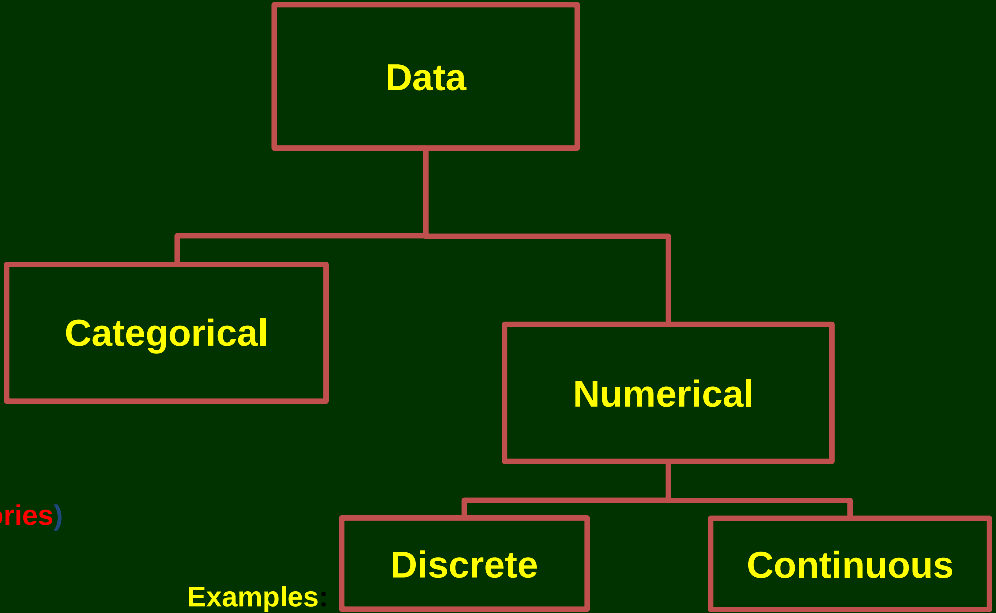
Business Domain Knowledge: Data analysts are more focused on understanding the business domain and the data they are analyzing, while data scientists require more in-depth knowledge of data science techniques, mathematics, and computer science.

Thus, data analysts focus on generating insights from data to support business decision-making, whereas data scientists are more focused on developing and implementing complex models and algorithms to solve more complex business problems.

Four different levels of Data

- Types of Variables
- Levels of Data Measurement
- Compare the four different levels of Data:
  Nominal
  Ordinal
  Interval and
  Ratio
- Usage Potential of Various Levels of Data
- Data Level, Operations, and Statistical Methods

Types of Variables



Data

Categorical

Numerical

Discrete

Continuous

Examples:

- Marital Status
- Political Party
- Eye Color

(Defined categories)

Examples:

- Number of Children
- Defects per hour

(Counted items)

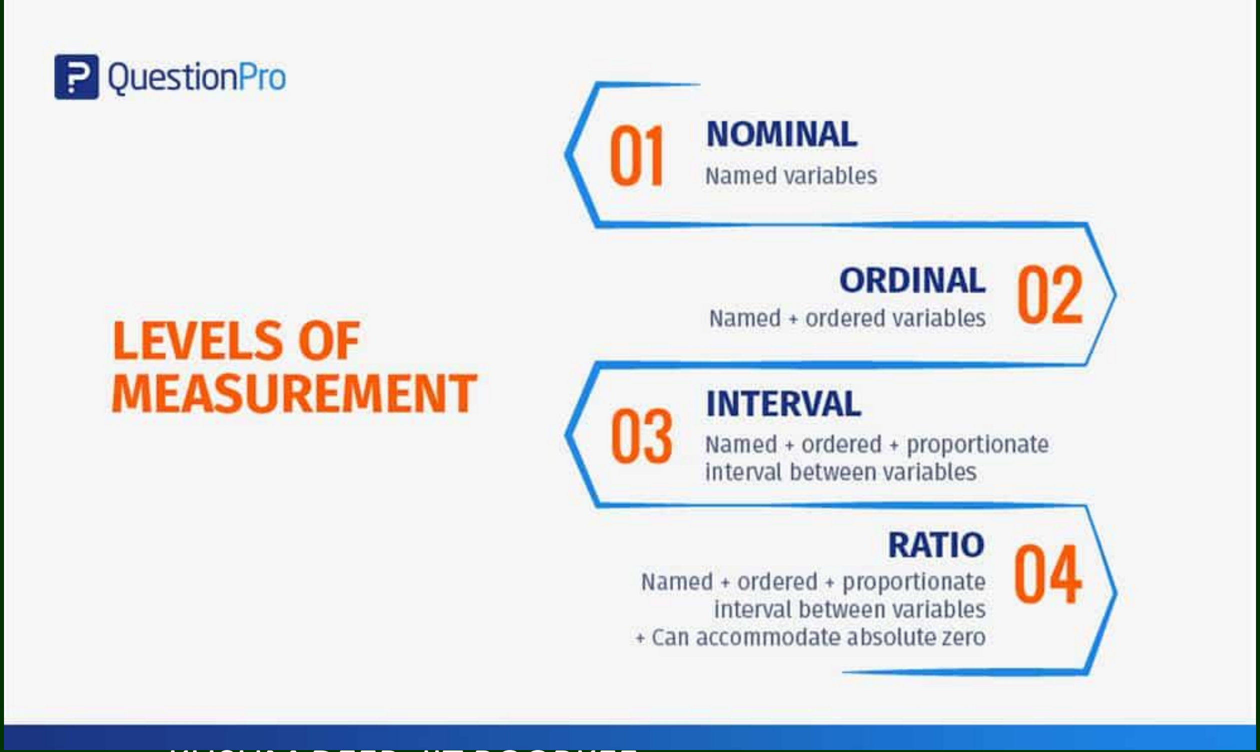
Examples:

- Weight
- Voltage

(Measured characteristics)

Levels of Data Measurement

1. Nominal
2. Ordinal
3. Interval
4. Ratio



QuestionPro

LEVELS OF MEASUREMENT

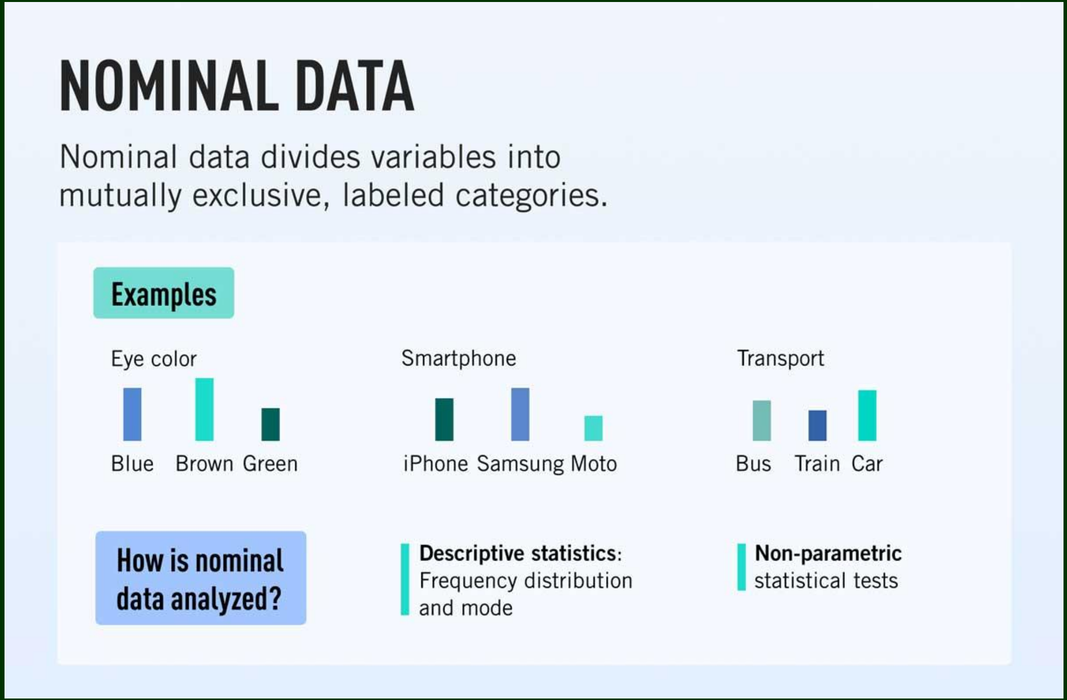
01 NOMINAL
Named variables

02 ORDINAL
Named + ordered variables

03 INTERVAL
Named + ordered + proportionate interval between variables

04 RATIO
Named + ordered + proportionate interval between variables + Can accommodate absolute zero

Examples of nominal data



NOMINAL DATA

Nominal data divides variables into mutually exclusive, labeled categories.

Examples

Eye color
Blue
Brown
Green
Smartphone
iPhone
Samsung
Moto
Transport
Bus
Train
Car

How is nominal data analyzed?

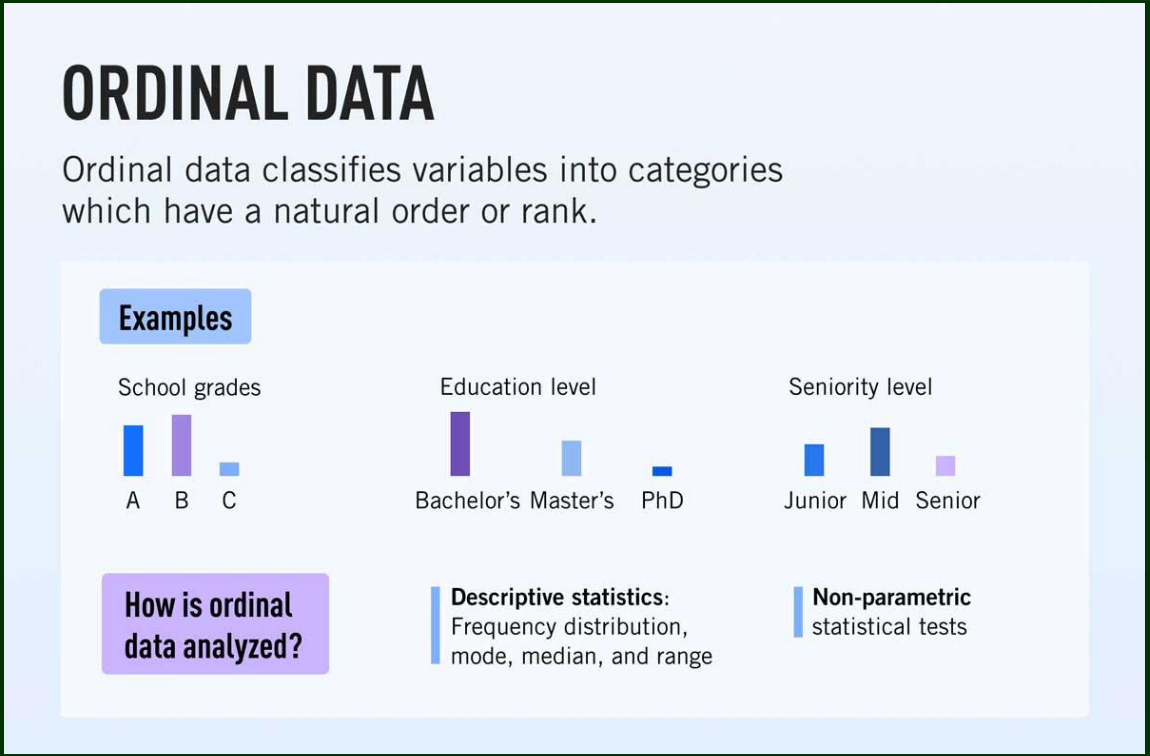
Descriptive statistics: Frequency distribution and mode

Non-parametric statistical tests

Give 5 examples of nominal data

1. Type of vehicle: Car, Truck, SUV, Van, Motorcycle
2. Political party affiliation: Republican, Democrat, Independent, Other
3. Type of housing: Apartment, Condo, Townhouse, Single-family home, Mobile home
4. Level of education: High School, Bachelor's degree, Master's degree, PhD, Other
5. Religion: Christianity, Islam, Judaism, Hinduism, Buddhism, Other
6. Language spoken: English, Spanish, French, German, Mandarin, Other
7. Type of fruit: Apple, Banana, Orange, Mango, Pineapple, Other
8. Country of origin: United States, Canada, Mexico, Japan, Brazil, Other
9. Mode of transportation: Walk, Bike, Car, Bus, Train, Other
10. Type of music: Rock, Pop, Hip-hop, Country, Jazz, Other

Examples of ordinal data



ORDINAL DATA

Ordinal data classifies variables into categories which have a natural order or rank.

Examples

School grades

A B C

Education level

Bachelor's Master's PhD

Seniority level

Junior Mid Senior

How is ordinal data analyzed?

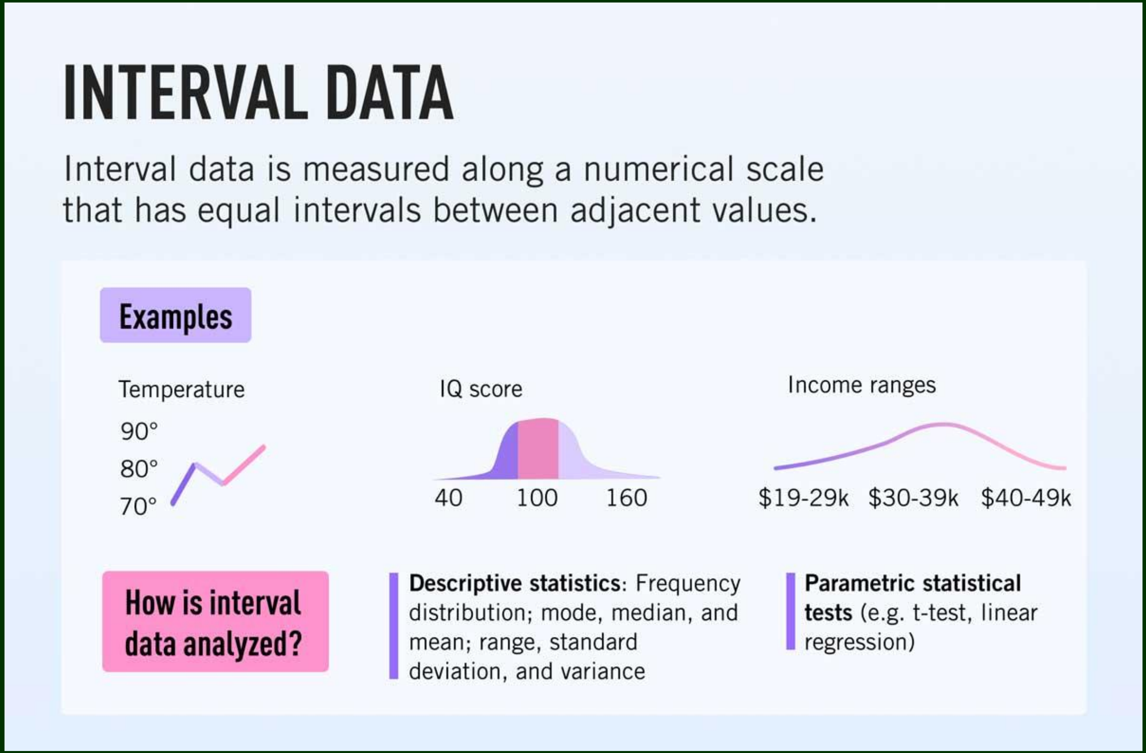
Descriptive statistics: Frequency distribution, mode, median, and range

Non-parametric statistical tests

Give 5 examples of ordinal data

1. Educational level: Elementary, Middle, High School, Associate's degree, Bachelor's degree, Master's degree, PhD
2. Economic status: Poor, Lower middle class, Middle class, Upper middle class, Rich
3. Job experience: Entry-level, Junior, Senior, Manager, Executive
4. Likert scale ratings: Strongly disagree, Disagree, Neutral, Agree, Strongly agree
5. Performance ratings: Poor, Fair, Good, Very good, Excellent
6. Customer satisfaction ratings: Very dissatisfied, Dissatisfied, Neutral, Satisfied, Very satisfied
7. Education grades: A, B, C, D, F
8. Pain level: No pain, Mild pain, Moderate pain, Severe pain, Extreme pain
9. Levels of agreement: Strongly disagree, Disagree, Neither agree nor disagree, Agree, Strongly agree
10. Stage of cancer: Stage 0, Stage 1, Stage 2, Stage 3, Stage 4.

Examples of interval data



INTERVAL DATA

Interval data is measured along a numerical scale that has equal intervals between adjacent values.

Examples

Temperature
90°
80°
70°
IQ score
40
100
160
Income ranges
$19-29k
$30-39k
$40-49k

How is interval data analyzed?

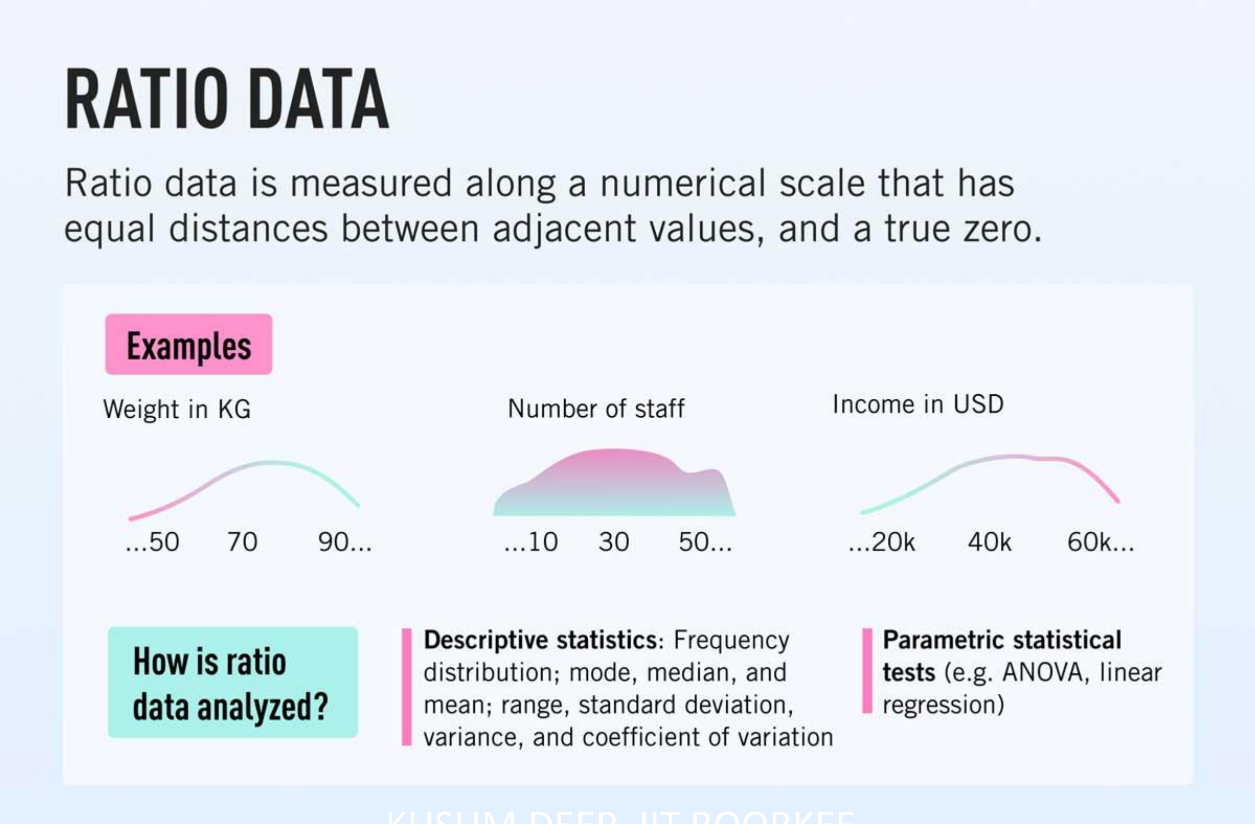
Descriptive statistics: Frequency distribution; mode, median, and mean; range, standard deviation, and variance

Parametric statistical tests (e.g. t-test, linear regression)

5 examples of interval data

1. Temperature in Celsius or Fahrenheit
2. Time of the day in hours, minutes, and seconds
3. Dates measured in days, months, or years
4. Standardized test scores, such as SAT or GRE
5. IQ scores
6. Credit scores
7. pH level of a solution
8. Altitude in feet or meters
9. Blood pressure readings
10. Annual income in dollars.

Examples of ratio data



RATIO DATA

Ratio data is measured along a numerical scale that has equal distances between adjacent values, and a true zero.

Examples

Weight in KG
...50
70
90...
Number of staff
...10
30
50...
Income in USD
...20k
40k
60k...

How is ratio data analyzed?

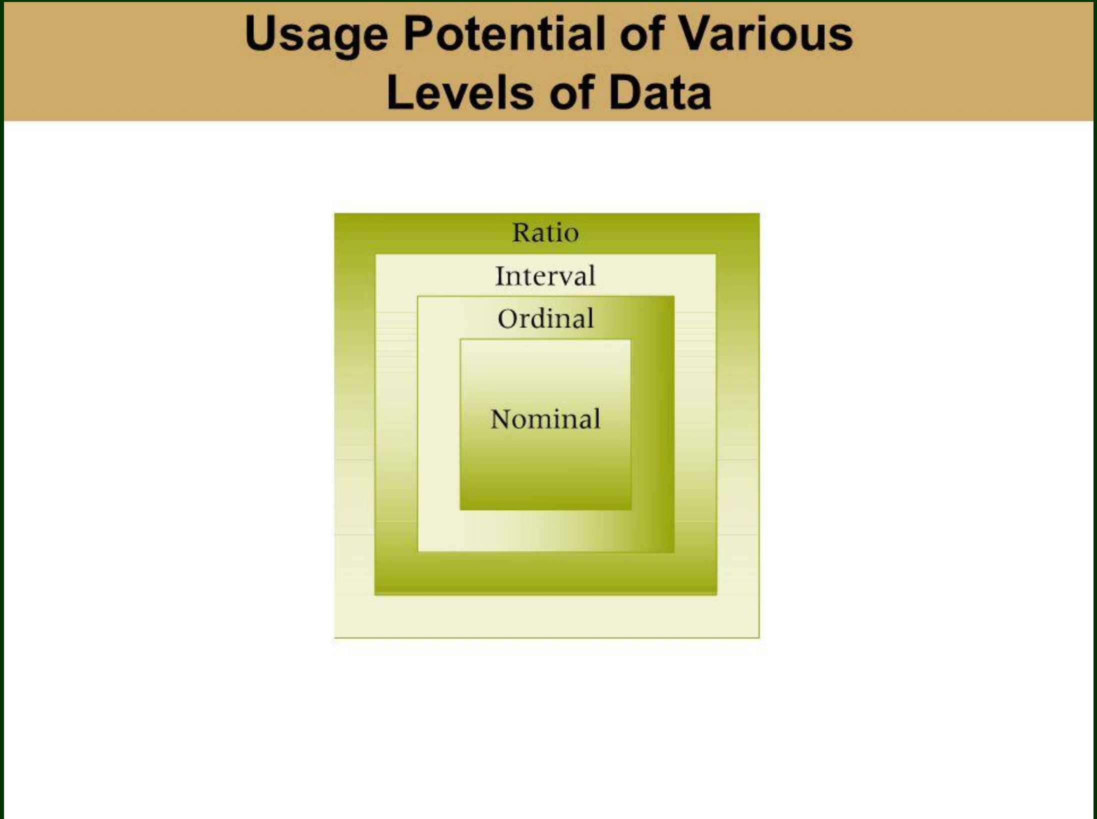
Descriptive statistics: Frequency distribution; mode, median, and mean; range, standard deviation, variance, and coefficient of variation

Parametric statistical tests (e.g. ANOVA, linear regression)

Give 5 examples of Ratio data

1. Age in years
2. Reaction time in milliseconds
3. Force in Newtons
4. Power in Watts
5. Velocity in meters per second
6. Acceleration in meters per second squared
7. Electric current in amperes
8. Resistance in ohms
9. Heart rate in beats per minute
10. Volume in liters or gallons

Usage Potential of Various Levels of Data



Usage Potential of Various Levels of Data

Impact of choice of measurement scale

Data Level	Meaningful Operations	Statistical Methods
Nominal	Classifying and Counting	Nonparametric
Ordinal	All of the above plus Ranking	Nonparametric
Interval	All of the above plus Addition, Subtraction	Parametric
Ratio	All of the above plus multiplication and division	Parametric

Case Study

Case Study: Fraud Detection in Credit Card Transactions

Background:

A financial institution wants to reduce the risk of fraudulent credit card transactions. They want to analyze their transaction data to identify patterns that may indicate fraudulent activity and develop strategies to prevent it.

Objective:

To use data analytics to identify patterns that indicate fraudulent credit card transactions and develop strategies to prevent it.

Methodology:

Data Collection: The financial institution collects data from credit card transactions, including transaction amount, location, time, and other relevant information. This data is then cleaned and prepared for analysis.

Exploratory Data Analysis: The data is analyzed to identify patterns, correlations, and trends. This helps to identify the factors that are most strongly associated with fraudulent transactions, such as unusually large transactions, transactions in different countries, and transactions that occur at unusual times.

Machine Learning Model Building: The data is split into training and testing sets, and a machine learning model is built to predict fraudulent transactions. The model uses various features, such as transaction amount, location, and time, to predict whether a transaction is likely to be fraudulent or not.

Model Evaluation: The model is evaluated using metrics such as accuracy, precision, recall, and F1-score. This helps to determine the effectiveness of the model and identify areas for improvement.

Strategy Development: Based on the insights gained from the analysis, the financial institution develops strategies to prevent fraudulent transactions. These may include monitoring transactions in real-time, flagging suspicious transactions for review, or blocking transactions that are deemed high risk.

Results:

The data analysis revealed that unusually large transactions, transactions in different countries, and transactions that occur at unusual times were the most significant factors contributing to fraudulent transactions. The machine learning model achieved an accuracy of 90%, which was deemed satisfactory. The financial institution used the insights gained from the analysis to develop strategies to prevent fraudulent transactions, including real-time transaction monitoring and flagging suspicious transactions for review. This resulted in a 25% reduction in fraudulent transactions over the next quarter.

Conclusion:

Data analytics can help financial institutions identify patterns that indicate fraudulent credit card transactions and develop strategies to prevent it. By using machine learning models to predict fraudulent transactions, financial institutions can monitor transactions in real-time, flag suspicious transactions for review, and block high-risk transactions to reduce the risk of fraudulent activity and protect their customers.

Steps in Exploratory Data Analysis

Def: **Exploratory Data Analysis (EDA)** is the process of examining and analyzing data sets to summarize their main characteristics and extract valuable insights.

Steps in conducting an EDA are:

1. **Question definition**: Identifying weakness and define objectives.
2. **Data Collection:** Collecting the relevant data from various sources, such as databases, spreadsheets, and other data repositories.
3. **Data Cleaning:** Pre-processing the data to ensure that it is consistent, complete, and accurate. This step involves checking for missing values, duplicates, outliers, and other anomalies in the data.

4. **Data Exploration:** Performing basic statistical analyses, such as calculating summary statistics, frequency distributions, and visualizations, to gain an initial understanding of the data.
5. **Data Visualization:** Creating various graphical representations, such as scatter plots, histograms, and heat maps, to visualize the data and identify patterns and trends.
6. **Feature Engineering:** Extracting and creating new features from the existing data, such as calculating ratios or transforming variables, to improve the predictive power of the model.

7. **Data Modeling:** Building statistical or machine learning models to predict or classify the data based on the patterns and relationships identified in the previous steps.
8. **Model Evaluation:** Evaluating the performance of the model using various metrics, such as accuracy, precision, recall, and F1-score, and comparing it with other models or benchmarks.
9. **Communication:** Presenting the findings and insights obtained from the EDA process in a clear and concise manner using reports, dashboards, or presentations, to facilitate decision-making and enhance understanding of the data.

Step 1: Question Definition – examples

- Which category of customers are not loyal to us any longer?
- What is the distribution of the data?
- Are there any outliers or extreme values?
- What is the central tendency of the data?
- What is the spread or variability of the data?
- Are there any relationships or patterns in the data?
- How do different groups within the data compare?
- What are the most important variables that explain the variation in the data?

Step 2: Data Collection

Three categories of data:

1. **Primary data**: Data collected by the analyst / the company directly
2. **Secondary data**: Data acquired from other organizations
3. **Third–party data**: Data collected and organized from numerous sources.

Step 2: Data Collection - examples

Surveys: can be conducted in person, by phone, or online and can collect both quantitative and qualitative data.

Observation: Observational studies involve watching and recording data on a particular behavior or event. Usually used in fields such as psychology or anthropology.

Secondary data: Secondary data involves collecting data that has already been collected and analyzed by someone else. Examples include data from government agencies or previous research studies.

Interviews: can be conducted in person, by phone, or online and can collect qualitative data on a particular topic or issue.

Experiments: Involve manipulating one or more variables to observe their effect on the outcome variable. Useful in fields such as psychology or biology.

Focus groups: Focus groups involve bringing together a small group of people to discuss a particular topic or issue. Useful in marketing research.

Example

Suppose a company wants to collect data on its employees' job satisfaction, job performance, and demographics. The company decides to conduct a survey that includes the following questions:

On a scale of 1 to 10, how satisfied are you with your job?

On a scale of 1 to 10, how would you rate your job performance?

What is your age?

What is your gender?

What is your job title?

How long have you been with the company?

The company sends the survey to all its employees and receives the following responses from a sample of 10 employees:

Employee	Job Satisfaction	Job Performance	Age	Gender	Job Title	Tenure
1	7	8	32	Male	Manager	5
2	6	6	27	Female	Analyst	1
3	8	9	41	Male	Director	10
4	5	4	45	Female	Manager	3
5	9	9	29	Male	Analyst	2
6	6	7	35	Female	Director	8
7	4	3	22	Male	Intern	0
8	7	8	38	Female	Manager	6
9	9	9	44	Male	Director	11
10	6	5	31	Female	Analyst	4

In this example, the company has collected multidimensional data that includes numerical (job satisfaction, job performance, age, tenure) and categorical (gender, job title) variables for each employee in the sample. The data was collected using a survey that includes questions about different aspects of the employees' job satisfaction, performance, and demographics. The company can use this data to analyze the relationships between job satisfaction, job performance, and other variables, such as age, gender, job title, and tenure, and to identify factors that contribute to employee satisfaction and performance.

Steps 3: Data Cleaning – what it is

It involves identifying and correcting errors, inconsistencies, and inaccuracies in the data. It is necessary because data can be collected from multiple sources and can be incomplete, inconsistent, or contain errors. These are the steps:

Data audit: The data is audited to identify any inconsistencies, missing data, or errors.

Data profiling: It involves analyzing the data to identify any patterns, trends, or anomalies.

Data validation: Involves verifying the accuracy of the data by comparing it with known constraints or business rules.

Data transformation: Data transformation involves converting the data into a standardized format, such as converting units of measurement or reformatting dates.

Data imputation: Involves filling in missing values with plausible values using statistical techniques.

Data merging: Involves combining data from multiple sources into a single dataset.

Data deduplication: Involves removing any duplicate data points from the dataset.

Steps 3: Data Cleaning – examples

Removing duplicates: e.g.

Multiple entries of same customer in a customer database due to data entry errors.

Duplicate transactions in a financial dataset that occur due to technical glitches or software errors.

Multiple copies of same email in an email database due to accidental forwarding or copying.

Multiple identical survey responses from the same respondent due to technical issues or respondent error.

Repetitive sensor readings in a sensor dataset due to sensor malfunction or incorrect data recording.

Example

Suppose a researcher has collected data on the number of hours of sleep each participant gets per night. However, due to a technical error, some of the participants were accidentally entered twice in the dataset. Here is an example of the dataset:

Participant ID	Hours of Sleep
1	7
2	6
3	8
4	7
5	6
2	5
6	9
7	8
8	7
9	6
10	8
1	6

In this example, participants 2 and 1 are entered twice in the dataset, which could cause problems in data analysis. To remove the duplicates, the researcher can use data cleaning techniques such as sorting the data by participant ID and then removing the duplicate rows. After removing the duplicates, the dataset would look like this:

Earticipant ID	Hours of Sleep
1	7
2	6
3	8
4	7
5	6
6	9
7	8
8	7
9	6
10	8

Steps 3: Data Cleaning – examples

Handling missing data. Some techniques are:

- **Deleting missing data**
- **Imputing missing data**. e,.g. mean imputation, median imputation, mode imputation, regression imputation, and multiple imputation.
- **Using the previous or next value**: In time series data, missing values can be replaced with the previous or next value in the series, assuming the missing value is not an outlier.
- **Using external data**: If external data is available that is correlated with the missing data.

There are three categories of missing data:

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Not Missing at Random (NMAR)

Example of Missing Data

Suppose a researcher is conducting a study on the relationship between stress and blood pressure. The researcher collects data from 20 participants, but due to various reasons, some data is missing. Here is an example of the dataset:

Participant ID	Stress Level (0-10)	Blood Pressure (mmHg)
1	8	130
2	6	120
3	7	128
4	9	NaN
5	5	122
6	NaN	125
7	7	129
8	NaN	NaN
9	6	127
10	8	130
11	6	123
12	7	NaN
13	NaN	126
14	7	128
15	5	124
16	8	NaN
17	NaN	125
18	6	122
19	9	131
20	5	121

In this example, there are missing values (represented by "NaN") for both the stress level and blood pressure variables for some participants. The researcher needs to address these missing values before analyzing the data. One approach is to remove the rows with missing values, but this can lead to a loss of valuable information. Another approach is to impute the missing values, for example, by replacing the missing values with the mean or median of the variable.

In this case, the researcher decides to impute the missing values with the mean of the respective variable. After imputing the missing values, the dataset would look like this:

articipant ID	Stress Level (0-10)	Blood Pressure (mmHg)
1	8	130
2	6	120
3	7	128
4	9	126.05
5	5	122
6	6.9	125
7	7	129
8	6.9	125.75
9	6	127
10	8	130
11	6	123
12	7	126.05
13	6.9	126
14	7	128
15	5	124
16	8	126.05
17	6.9	125
18	6	122
19	9	131
20	5	121

Steps 3: Data Cleaning – examples

Examples of handling missing data are:

- **Regression imputation**: Uses a regression model to predict the missing values based on the values of the other variables in the dataset. E.g. if a person's income is missing in a survey, it can be imputed using regression analysis with other variables, such as age, education, and occupation.

Example of Regression imputation

Suppose the dataset has two variables:
X and Y, where Y has some missing values.
First train a regression model using X as the predictor variable and Y as the response variable using linear regression model. Then, use the non-missing values of Y to train the model, and then use the model to predict the missing values of Y based on the corresponding values of X.

X	Y
10.0	25.0
5.0	10.0
2.0	NaN
8.0	20.0
6.0	NaN
7.0	18.0
4.0	NaN
3.0	8.0
1.0	NaN
9.0	22.0

```
import numpy as np
from sklearn.linear_model import LinearRegression
# Load the data
X = np.array([10.0, 5.0, 2.0, 8.0, 6.0, 7.0, 4.0, 3.0, 1.0, 9.0]).reshape(-1, 1)
Y = np.array([25.0, 10.0, np.nan, 20.0, np.nan, 18.0, np.nan, 8.0, np.nan, 22.0])
# Train the regression model
model = LinearRegression().fit(X[Y==Y], Y[Y==Y])
# Use the model to predict the missing values
Y_imputed = model.predict(X)
Y[np.isnan(Y)] = Y_imputed[np.isnan(Y)]
```

The missing values of Y have been imputed using the regression model, and the dataset now contains no missing values.

This is the cleaned data.

X	Y
10.0	25.0
5.0	10.0
2.0	5.5
8.0	20.0
6.0	15.5
7.0	18.0
4.0	8.5
3.0	8.0
1.0	2.0
9.0	22.0

Using external data: If the missing data is related to weather patterns, it can be imputed using weather data from a nearby weather station.

Ex: Let's consider a scenario where we have a dataset of 100 patients with missing values for their blood pressure readings. We want to impute these missing values using external data from a separate dataset of 500 patients with complete blood pressure readings. Here's how we can perform the imputation:

First, we need to understand the relationship between the variables in our two datasets. We can do this by calculating the correlation coefficient between the blood pressure readings in the two datasets. Let's say the correlation coefficient is 0.8, indicating a strong positive relationship.

Next, we can use a simple linear regression model to predict the missing blood pressure values in our dataset using the complete blood pressure readings from the external dataset. We'll use the blood pressure readings in the external dataset as the independent variable and the blood pressure readings in our dataset as the dependent variable.

Once we have our regression model, we can use it to predict the missing blood pressure values in our dataset. For example, let's say the model predicts that the missing blood pressure reading for patient 1 is 120/80 mmHg.

We repeat this process for all the missing values in our dataset until we have imputed all the missing values.

Finally, we can assess the accuracy of our imputation by comparing the imputed values to the true values (if available) or by examining the distribution of imputed values.

Multiple imputation: This is a more complex method that involves imputing multiple plausible values for each missing data point, resulting in a set of datasets that can be used for analysis.

Ex: Suppose we have a dataset with 500 observations on variables X, Y, and Z, but there are missing values in some of the observations for variables Y and Z.

Use multiple imputation to fill in the missing values.

First, we create multiple imputed datasets. Let's say we generate 5 imputed datasets, meaning we will fill in the missing values 5 different times to create 5 different complete datasets.

For each imputed dataset, we create a predictive model to estimate the missing values. In this example, let's say we create a regression model to predict the missing values of Y and Z based on X and other available variables.

We fit this model on the complete cases, i.e., observations that have values for all three variables X, Y, and Z. Then, we use the fitted model to predict the missing values in the incomplete cases, i.e., observations that have missing values for Y or Z.

We repeat this process for each imputed dataset, creating 5 different sets of imputed values for Y and Z.

After creating the 5 imputed datasets, we analyze each of them separately using the analysis method of our choice (e.g., regression analysis, clustering analysis, etc.) to obtain 5 different sets of results.

Finally, we combine the results from the 5 different imputed datasets using appropriate procedures for multiple imputation. For example, we can calculate the mean, variance, and standard error of the estimates from each imputed dataset to obtain the overall estimates and their standard errors.

Imputation Techniques

Mean imputation: Replaces missing values with the mean value of the non-missing values for that variable.

Median imputation: Replaces missing values with the median value of the non-missing values for that variable.

Mode imputation: Replaces missing values with the mode value (i.e., the most frequent value) of the non-missing values for that variable.

Ex: Suppose we have a dataset of 1000 observations on a variable X, but 100 of these observations have missing values. We want to fill in the missing values using three different methods: mean imputation, median imputation, and mode imputation.

Mean imputation: We calculate the mean of the non-missing values of X and replace the missing values with this mean value. Let's say the mean of the non-missing values is 10.5. We replace the missing values with 10.5 and obtain a new dataset with no missing values.

Median imputation: We calculate the median of the non-missing values of X and replace the missing values with this median value. Let's say the median of the non-missing values is 11. We replace the missing values with 11 and obtain a new dataset with no missing values.

Mode imputation: We calculate the mode of the non-missing values of X and replace the missing values with this mode value. Let's say the mode of the non-missing values is 12. We replace the missing values with 12 and obtain a new dataset with no missing values.

After performing the imputations, we can analyze the new datasets to assess the impact of each imputation method on our results. For example, we can calculate the mean, median, and mode of the complete datasets and compare them to the imputed values. We can also examine the distribution of the imputed values and check if there are any outliers or unusual patterns.

Note that mean, median, and mode imputation can introduce bias and reduce the variability of the data. Therefore, these methods should be used with caution and only when the missing values are believed to be missing at random and not related to the values of other variables. Additionally, mean and median imputation can only be used for numerical variables, while mode imputation can only be used for categorical variables.

Hot deck imputation: Replaces missing values with values from other similar records in the dataset. This is done by finding records with similar values for the other variables and using their values to fill in the missing values.

Ex: Suppose we have a dataset of 2000 observations on variables X and Y, but there are missing values in some of the observations for variable Y. We want to use hot deck imputation to fill in the missing values.

First, we sort the observations based on the values of variable X.

Then, we create groups or "decks" of observations with similar values of X. For example, we may create a deck for observations with X values between 1 and 10, another deck for X values between 11 and 20, and so on.

For each observation with a missing value for Y, we find the nearest neighbor in the same deck based on the values of X. This nearest neighbor is an observation in the same deck that has a non-missing value for Y and is most similar to the observation with the missing value.

We use the Y value from the nearest neighbor as the imputed value for the missing Y value in the observation with the missing value.

After performing the hot deck imputation, we have a new dataset with no missing values for Y.

Note that hot deck imputation can introduce bias if the nearest neighbor is not a good match for the observation with the missing value. Additionally, hot deck imputation assumes that the values of Y are related to the values of X, which may not always be the case. Therefore, this method should be used with caution and the results should be carefully examined to assess their validity.

K-nearest neighbor imputation: Replaces missing values with the values of the K most similar records in the dataset, where similarity is based on the values of the other variables in the dataset.

Ex: Suppose we have a dataset of 500 observations on variables X, Y, and Z, but there are missing values in some of the observations for variable Y. We want to use KNN imputation to fill in the missing values.

First, we identify a distance metric to measure the similarity between observations. A common distance metric is Euclidean distance.

Then, we select a value of k, which represents the number of nearest neighbors we want to use to impute the missing values. Let's say we choose k = 5.

For each observation with a missing value for Y, we calculate the Euclidean distance between this observation and all other observations in the dataset that have non-missing values for Y.

We select the k observations that have the smallest distances to the observation with the missing value. These k observations are the "nearest neighbors" of the observation with the missing value.

We use the Y values from the k nearest neighbors to calculate an imputed value for the missing Y value in the observation with the missing value. One common method is to take the average of the Y values from the k nearest neighbors.

After performing the KNN imputation, we have a new dataset with no missing values for Y.

Note that KNN imputation assumes that observations that are close to each other in the feature space are similar to each other. Therefore, this method should be used with caution and the results should be carefully examined to assess their validity. Additionally, KNN imputation can introduce bias if the value of k is not appropriate for the dataset. Choosing the optimal value of k is often done through trial and error or using cross-validation techniques.

Steps 3: Data Cleaning – examples

Examples of Handling outliers

Removing outliers: If a dataset of exam scores includes a score of 1000, which is much higher than any other score, it can be considered an outlier and removed from the dataset.

Transforming data: If a dataset of income includes a few very high salaries that are outliers, the data can be transformed using a logarithmic function to reduce the effect of outliers.

Winsorizing: If a dataset of housing prices includes a few very high prices that are outliers, the highest prices can be winsorized to the 95th percentile, meaning any prices above the 95th percentile are replaced with the 95th percentile value.

Using robust statistical methods: If a dataset of salaries includes a few very high salaries that are outliers, using a robust statistical method such as the median can reduce the effect of outliers.

Capping: If a dataset of electricity consumption includes a few very high values that are outliers, the values can be capped at a certain maximum value, such as the 99th percentile value, to reduce the effect of outliers.

Example of outliers

Suppose a researcher is collecting data on the salaries of employees at a company. The researcher collects data from 50 employees, and the salaries range from $40,000 to $200,000. However, the researcher notices that there are two values that are much higher than the rest of the salaries: $500,000 and $1,000,000. These two values are much larger than the other salaries, and are considered outliers.

An outlier is an observation that is significantly different from other observations in the dataset. Outliers can occur due to measurement errors, data entry errors, or due to the presence of extreme values in the population being sampled.

In this example, the researcher needs to decide whether to keep or remove the outliers from the dataset. If the outliers are due to data entry errors, the researcher may choose to remove them to avoid bias in the analysis. However, if the outliers are due to the presence of high salaries in the population being sampled, the researcher may choose to keep them in the analysis.

Steps 3: Data Cleaning – examples

Standardizing variables.

Standardizing height and weight: If a dataset includes measurements of height in centimeters and weight in kilograms, standardizing these variables can help to compare them. The variables can be standardized by converting height to meters and then dividing weight by the square of height in meters to calculate the body mass index (BMI).

Standardizing test scores: If a dataset includes test scores on different tests with different scales, standardizing the scores can help to compare them. The scores can be standardized by converting them to z-scores, which measure the number of standard deviations a score is from the mean.

Standardizing financial data: If a dataset includes financial data such as revenues and expenses that are measured in different currencies or scales, standardizing the data can help to compare them. The data can be standardized by converting them to a common currency or scale, such as converting all values to US dollars or dividing them by the total revenue.

Standardizing time-series data: If a dataset includes time-series data with different units of time, such as daily, weekly, or monthly data, standardizing the data can help to compare them. The data can be standardized by converting them to a common unit of time, such as daily data divided by seven for weekly data.

Steps 3: Data Cleaning – examples

Correcting data entry errors

Identifying typos and misspellings: If a dataset of customer names includes misspellings or typos, the errors can be corrected by checking the names against a list of correct names or using data cleaning software to identify and correct errors.

Handling missing data: If a dataset includes missing data, the missing values can be imputed or replaced using statistical methods or domain knowledge.

Correcting formatting errors: If a dataset of dates includes formatting errors, such as dates entered in the wrong format, the errors can be corrected by reformatting the dates using data cleaning software or by manually correcting the errors.

Addressing data duplication: If a dataset includes duplicated data due to errors in data entry, the duplicates can be removed or consolidated using data cleaning software or manual review.

Handling outliers: If a dataset includes values that are far outside the norm and are likely due to data entry errors, the values can be removed or corrected using data cleaning software or manual review.

Steps 3: Data Cleaning – examples

Handling inconsistent data: e.g. a person's age cannot be negative or exceed the maximum age limit.

Addressing different units of measurement: If a dataset includes measurements that are recorded in different units, such as kilometers and miles, the units can be converted to a common unit to facilitate comparison and analysis.

Correcting data range errors: If a dataset includes values that are outside the expected range for a given variable, such as negative values for age or weight, the errors can be corrected using data cleaning software or manual review.

Resolving conflicts between data sources: If a dataset is compiled from multiple sources and includes conflicting data, such as different birthdates for a single individual, the conflicts can be resolved by checking the accuracy of each data source and making corrections as needed.

Handling inconsistent data formats: If a dataset includes inconsistent data formats, such as dates entered in different formats or inconsistent capitalization, the data can be standardized using data cleaning software or manual review.

Dealing with incomplete or inconsistent data: If a dataset includes incomplete or inconsistent data, such as missing values or conflicting information, the data can be imputed or corrected using statistical methods or domain knowledge.

Step 4: Data Exploration – examples

Summary statistics: Calculate mean, median, mode, range, standard deviation, and correlation coefficients.

Histograms: This graphical representation of the frequency distribution of a variable shows shape of the distribution, the central tendency, and the variability of the data.

Box plots: This graphical representation shows the median, quartiles, range, and outliers.

Scatter plots: This graphical representation of the relationship between two variables, which shows the direction and strength of the relationship, as well as any outliers or clusters of points.

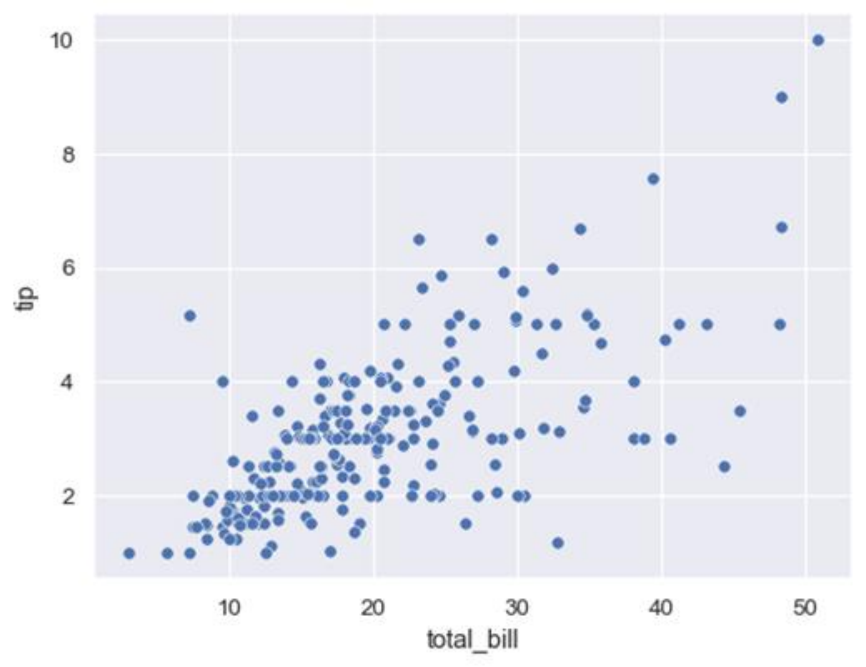
Heatmaps: This graphical representation of the correlation matrix of the variables, which shows the strength and direction of the correlation between pairs of variables.

Dimensionality reduction: This includes techniques such as principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE), which help to visualize the relationships between variables in a lower-dimensional space.

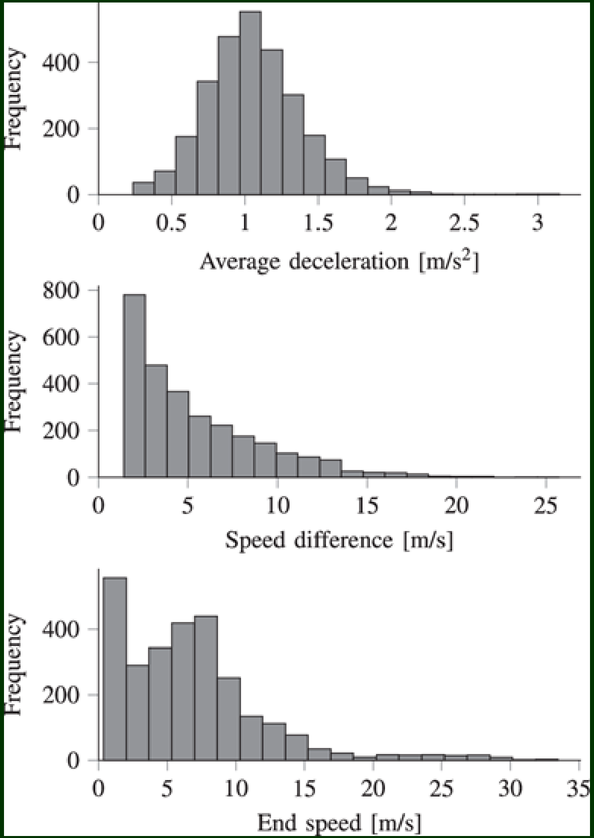
Step 5: Data Visualization - examples

Examples:

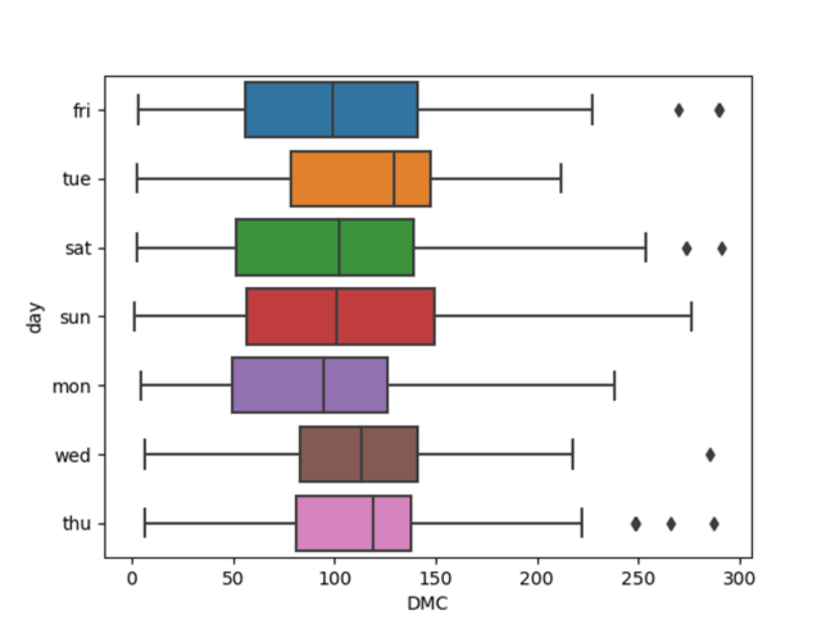
- **Scatter plots**



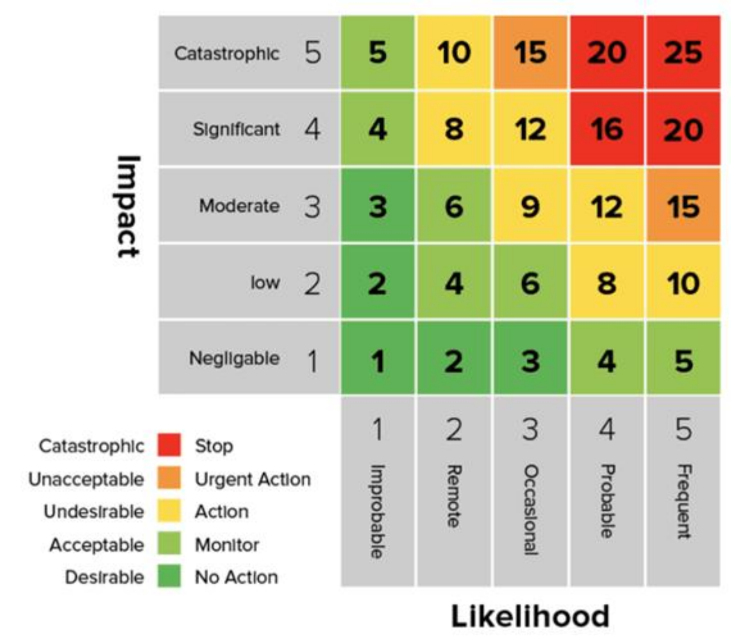
- Histograms


Frequency
400
200
0
0
0.5
1
1.5
2
2.5
3
Average deceleration [m/s^2]
Frequency
800
600
400
200
0
0
5
10
15
20
25
Speed difference [m/s]
Frequency
400
200
0
0
5
10
15
20
25
30
35
End speed [m/s]

- Box plots



- Heatmaps



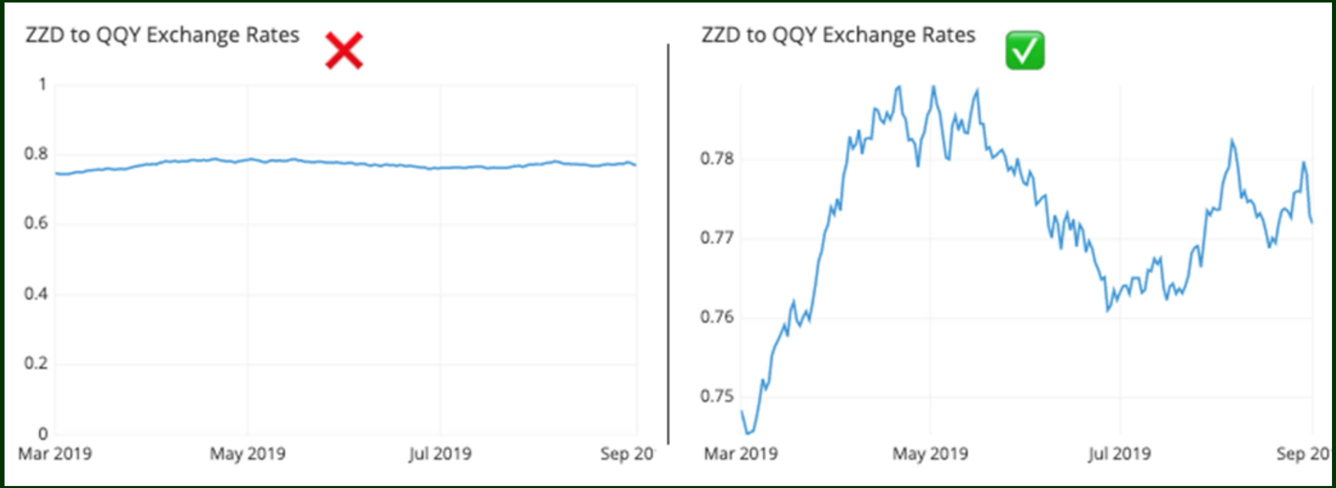
Impact

Catastrophic	5	5	10	15	20	25
Significant	4	4	8	12	16	20
Moderate	3	3	6	9	12	15
low	2	2	4	6	8	10
Negligable	1	1	2	3	4	5
		1 Improbable	2 Remote	3 Occasional	4 Probable	5 Frequent

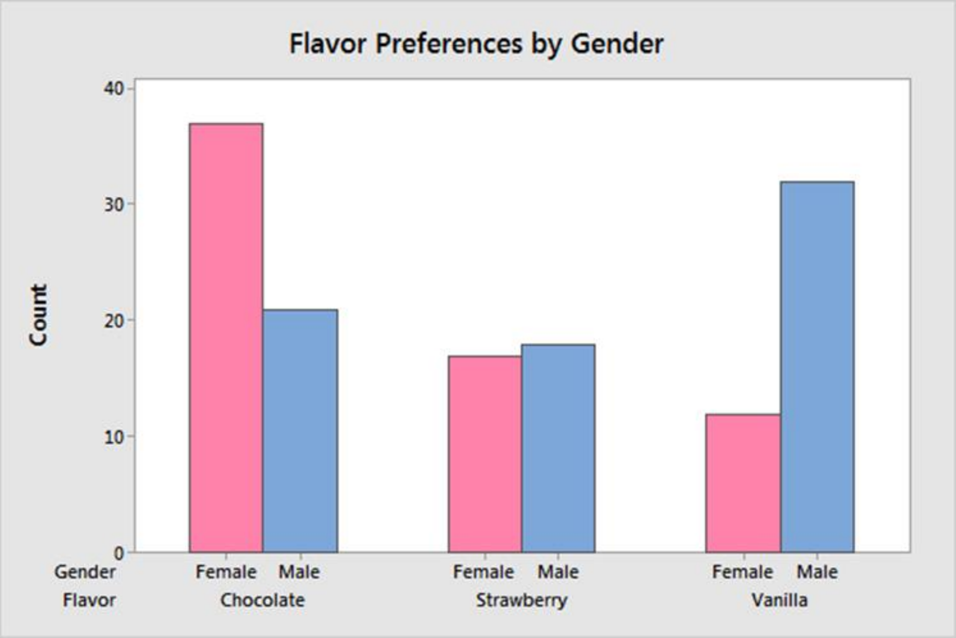
Likelihood

Catastrophic	🟥	Stop
Unacceptable	🟧	Urgent Action
Undesirable	🟨	Action
Acceptable	🟩	Monitor
Desirable	🟩	No Action

- Line plots


ZZD to QQY Exchange Rates
✗
1
0.8
0.6
0.4
0.2
0
Mar 2019
May 2019
Jul 2019
Sep 20
ZZD to QQY Exchange Rates
✓
0.78
0.77
0.76
0.75
Mar 2019
May 2019
Jul 2019
Sep 20

- Bar charts


Flavor Preferences by Gender
Count
0
10
20
30
40
Gender
Flavor
Female
Male
Chocolate
Female
Male
Strawberry
Female
Male
Vanilla

Principal Component Analysis

Suppose we have a dataset containing the following measurements on 5 different fruits:

	Apple	Banana	Orange	Peach	Pineapple
Size	10	15	8	12	20
Weight	4	6	3	5	8
Color	1	2	3	2	1

Here, each row represents a different fruit and each column represents a different measurement (size, weight, color). We can perform PCA on this dataset to identify the most important features that differentiate the fruits from each other.

First, we standardize the data by subtracting the mean of each measurement and dividing by the standard deviation. This gives us the following standardized dataset:

	Apple	Banana	Orange	Peach	Pineapple
Size	-0.67	0.00	-1.00	-0.33	2.00
Weight	-0.67	0.00	-1.00	-0.33	2.00
Color	-1.22	0.00	1.22	0.00	-1.22

Covariance matrix of the standardized dataset is:

	Size	Weight	Color
Size	1.00	1.00	-0.82
Weight	1.00	1.00	-0.82
Color	-0.82	-0.82	1.00

The eigenvectors and eigenvalues of the covariance matrix are

	Eigenvector	Eigenvalue
PC1	[-0.71, -0.71, 0.00]	2.44
PC2	[0.00, 0.00, -1.00]	0.18
PC3	[-0.71, 0.71, 0.00]	0.38

We can see that the first principal component (PC1) explains the most variance in the data (2.44 out of 3 total variance), while the second and third principal components (PC2 and PC3) explain relatively little variance.

Finally, we can project the data onto the first two principal components to visualize the fruits in a two-dimensional space:

	PC1	PC2
Apple	-1.03	0.00
Banana	-0.34	0.00
Orange	1.17	0.00
Peach	-0.34	0.00
Pineapple	2.54	0.00

We can see that the fruits are separated along the first principal component (which represents the size and weight of the fruits) but not along the second principal component (which represents the color of the fruits). This suggests that size and weight are the most important features that differentiate the fruits from each other.

Step 6: Feature Engineering - examples

It is the process of transforming raw data into features that can be used as input to machine learning algorithms. Some examples:

One-hot encoding: is a technique used to convert categorical variables into binary features that can be used in machine learning algorithms. Each category is represented as a binary feature, with a value of 1 indicating the presence of the category and 0 indicating its absence.

Example of One-hot encoding

Suppose we have a dataset containing information about 5 different people:

	Gender	Occupation
Person1	Male	Engineer
Person2	Female	Doctor
Person3	Male	Teacher
Person4	Female	Engineer
Person5	Male	Doctor

To perform one-hot encoding on this dataset, we create a new binary column for each possible value of each categorical variable. In this case, there are two possible values for the gender variable (Male and Female) and three possible values for the occupation variable (Engineer, Doctor, and Teacher). Therefore, we create a total of 5 new columns:

	Gen_M	Gen_F	Occ_Eng	Occ_Doct	Occ_Teacher
Person1	1	0	1	0	0
Person2	0	1	0	1	0
Person3	1	0	0	0	1
Person4	0	1	1	0	0
Person5	1	0	0	1	0

In each new column, a value of 1 indicates that the original variable had that value for that observation, and a value of 0 indicates that it did not. For example, in the first row of the table, Person1 is male and an engineer, so the Gender_Male and Occupation_Engineer columns have values of 1, while the other columns have values of 0.

One-hot encoding allows us to represent categorical variables as numerical variables that can be used in machine learning models. It also avoids the issue of ordinality that arises when encoding categorical variables using integer labels.

Binning: is a technique used to convert continuous variables into categorical features. The range of the variable is divided into a set of bins or intervals, and each value is assigned to the bin that it falls into.

Ex: Suppose we have a dataset of 1000 observations on a variable X, which represents the age of customers. We want to create a new feature based on X that groups the ages into different bins or categories.

First, we decide on the number and size of the bins we want to use. Let's say we want to use 4 bins: 0-24, 25-44, 45-64, and 65 and above.

Then, we create a new feature called "Age Group" that assigns each observation to one of the four bins based on the value of X. For example, an observation with X=30 would be assigned to the "25-44" bin.

We can now analyze the data using the new "Age Group" feature instead of the original "Age" feature. For example, we can calculate the mean or median of a target variable for each age group to see if there are any differences in behavior or preferences between the different age groups.

Binning can be useful for a number of reasons. For example, it can help to reduce the noise in the data and make it easier to spot trends and patterns. It can also help to deal with outliers and missing values, as observations can be assigned to a bin even if their value is missing or extreme. However, it's important to choose an appropriate number and size of bins to avoid overfitting or underfitting the data. Additionally, binning can result in loss of information, as the precise values of X are no longer used in the analysis.

Feature scaling: is a technique used to rescale the values of a variable to a common scale. This can help improve the performance of machine learning algorithms that are sensitive to the scale of the input features.

Ex: Suppose we have a dataset of 1000 observations on two variables X and Y, where X represents the income of customers in dollars and Y represents the age of customers in years. We want to use these variables to build a machine learning model to predict whether a customer is likely to make a purchase or not.

First, we examine the distribution of the variables to see if they need to be scaled. We notice that the income variable ranges from $10,000 to $100,000, while the age variable ranges from 18 to 80 years. Since the two variables have different scales, we decide to scale them before building the model.

One common method of scaling is standardization, which transforms the data to have a mean of 0 and a standard deviation of 1. We apply standardization to both variables by subtracting the mean and dividing by the standard deviation.

After scaling, the income variable now ranges from -3.16 to 3.16 and the age variable ranges from -2.56 to 2.56. The two variables are now on the same scale and can be compared and used in the model without one variable dominating the other.

We can now build a machine learning model using the scaled variables as features. For example, we may use logistic regression or decision trees to predict the likelihood of a customer making a purchase based on their income and age.

Feature scaling is important because many machine learning algorithms are sensitive to the scale of the input features. Scaling can help to improve the performance and accuracy of the model by ensuring that all features are treated equally and not dominated by one another. However, it's important to choose an appropriate scaling method and to ensure that the scaling does not result in loss of information or introduce bias into the analysis.

Feature extraction: is technique used to transform raw data into a set of features that can be used in machine learning algorithms. This can involve extracting features such as frequency, duration, or intensity from raw data such as audio or image signals.

Ex: Suppose we have a dataset of 1000 observations on a variable X, which represents text reviews of a product. We want to use these reviews to build a machine learning model to predict whether a customer is likely to recommend the product or not.

First, we examine the reviews and notice that they contain a lot of information, including words, phrases, and emotions. To make the data more manageable and relevant for our analysis, we decide to extract features from the text reviews.

One common method of feature extraction is to use a technique called bag-of-words, which involves converting the text reviews into a matrix of word counts. We start by tokenizing the text reviews into individual words, removing stop words (common words like "the" and "and"), and stemming the remaining words (reducing them to their root form).

After preprocessing the text reviews, we create a matrix where each row corresponds to a review and each column corresponds to a word. The values in the matrix are the counts of each word in the corresponding review. For example, if the word "excellent" appears 3 times in a review, the corresponding value in the matrix would be 3.

We can now use this matrix as the input features for a machine learning model. For example, we may use logistic regression or decision trees to predict whether a customer is likely to recommend the product based on the presence or absence of certain words in their review.

Feature extraction is important because it can help to reduce the dimensionality of the data, make the data more relevant to the problem at hand, and improve the performance and accuracy of the model. However, it's important to choose an appropriate feature extraction technique and to ensure that the extracted features are meaningful and not redundant. Additionally, feature extraction can result in loss of information, as some information from the original data may not be captured in the extracted features.

Dimensionality reduction: is a techniques such as principal component analysis (PCA) or t-distributed stochastic neighbor embedding (t-SNE) can be used to reduce the dimensionality of high-dimensional data while preserving its important features.

Ex: Suppose we have a dataset of 1000 observations on 50 variables, where each variable represents a different aspect of customer behavior on an e-commerce website. We want to use these variables to build a machine learning model to predict whether a customer is likely to make a purchase or not.

First, we examine the variables and notice that some of them may be highly correlated or redundant, meaning that they provide similar information. To make the data more manageable and to reduce the risk of overfitting, we decide to perform dimensionality reduction.

One common method of dimensionality reduction is to use a technique called principal component analysis (PCA), which involves transforming the data into a new set of variables called principal components. These principal components are linear combinations of the original variables that capture the maximum amount of variance in the data.

After applying PCA, we obtain a new set of variables that are uncorrelated with each other and explain the maximum amount of variance in the original data. For example, we may find that the first principal component is a combination of variables related to browsing behavior on the website, while the second principal component is a combination of variables related to purchase history.

We can now use these principal components as the input features for a machine learning model. For example, we may use logistic regression or decision trees to predict whether a customer is likely to make a purchase based on their behavior on the website, as captured by the principal components.

Dimensionality reduction is important because it can help to reduce the complexity of the data, make the data more interpretable, and improve the performance and accuracy of the model. However, it's important to choose an appropriate dimensionality reduction technique and to ensure that the reduced data still captures the relevant information from the original data. Additionally, dimensionality reduction can result in loss of information, as some information from the original data may not be captured in the reduced data.

Step 7: Data Modelling

It is a process of creating a conceptual representation of data and defining its structure, relationships, and constraints, so as to understand underlying patterns and relationships in the data. Some examples:

Entity-relationship (ER) modeling: is a technique used to represent the relationships between different entities in a dataset to identify key features and relationships within the data. E.g. in a customer transaction dataset, it can be be used to identify the relationships between customers, transactions, and products.

Ex: ER Modelling : The Flight Database

The flight database stores details about an airline’s fleet, flights, and seat bookings. It’s a hugely simplified version of what a real airline would use, but the principles are the same. Consider the following requirements list:

The airline has one or more airplanes.

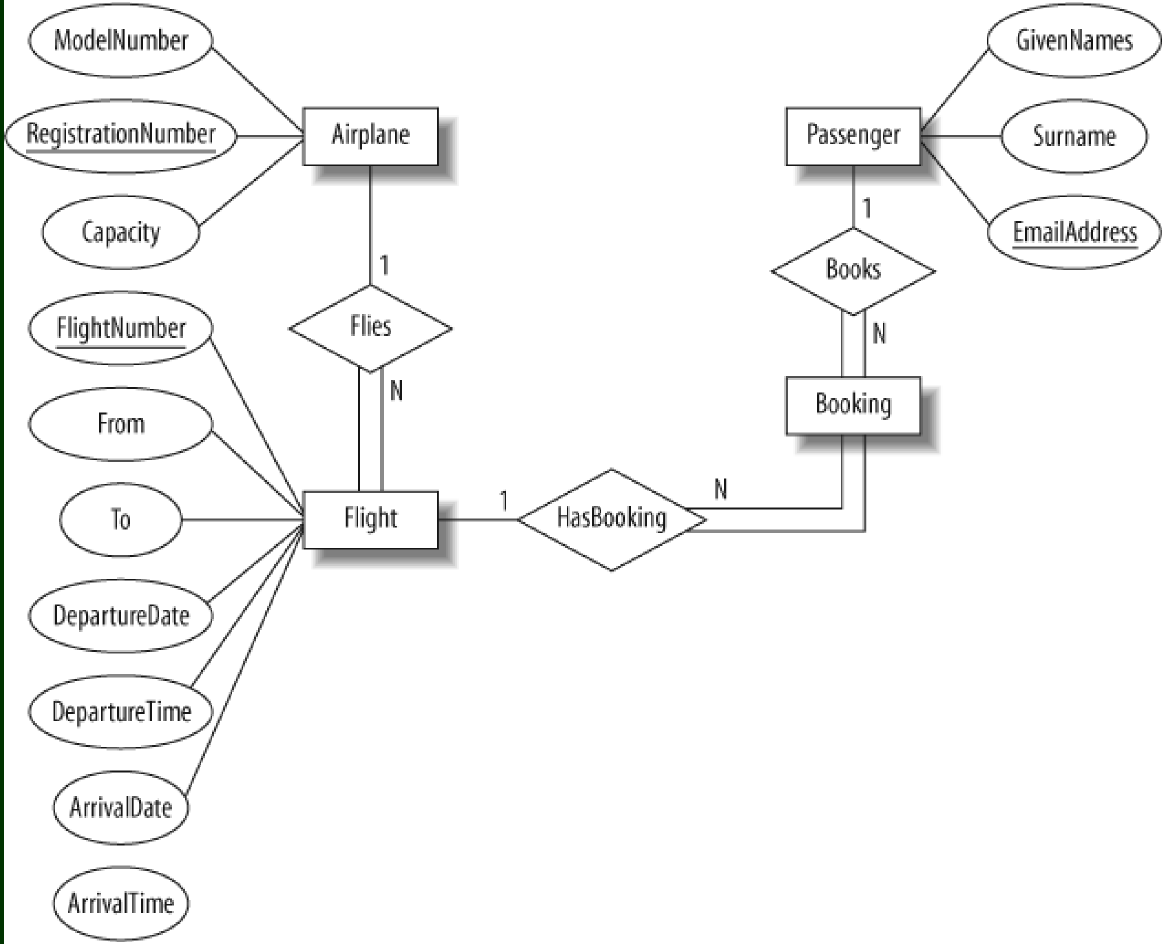
An airplane has a model number, a unique registration number, and the capacity to take one or more passengers.

An airplane flight has a unique flight number, a departure airport, a destination airport, a departure date and time, and an arrival date /time.

Each flight is carried out by a single airplane.

A passenger has given names, a surname, and a unique email address.

A passenger can book a seat on a flight.


ModelNumber
RegistrationNumber
Capacity
Airplane
1
Flies
N
FlightNumber
From
To
DepartureDate
DepartureTime
ArrivalDate
ArrivalTime
Flight
1
HasBooking
N
Booking
N
Books
1
Passenger
GivenNames
Surname
EmailAddress

Time-series modeling: is used to analyze time-series data, so as to identify trends and patterns in the data. E.g. in a financial dataset, it can be used to analyze stock prices over time and identify trends or patterns.

Ex: Suppose we have monthly sales data for a product over the past 24 months, as follows:

Month	Sales
Jan-21	100
Feb-21	120
Mar-21	150
Apr-21	130
May-21	140
Jun-21	160
Jul-21	180
Aug-21	200
Sep-21	190
Oct-21	210
Nov-21	220
Dec-21	240
Jan-22	110
Feb-22	130
Mar-22	160
Apr-22	140
May-22	150
Jun-22	170
Jul-22	190
Aug-22	220
Sep-22	200
Oct-22	230
Nov-22	240
Dec-22	260

We want to forecast the sales for the next 6 months (Jan-23 to Jun-23).

We can use ARIMA, a commonly used time series modeling technique, to forecast the sales. ARIMA stands for Autoregressive Integrated Moving Average, and it is a statistical method that models the dependence between an observation and a number of lagged observations and forecast errors.

First, we need to check if the time series is stationary, which means that the mean, variance, and autocorrelation structure of the series do not change over time. We can use the Augmented Dickey-Fuller (ADF) test for this. If the series is not stationary, we can difference it to make it stationary.

After we have a stationary series, we can estimate the parameters of the ARIMA model using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to select the optimal model. In this example, we will assume that the optimal model is ARIMA(1,1,1).

We can then use the ARIMA model to forecast the sales for the next 6 months. The results of the forecast are as follows:

These forecasts can help businesses make informed decisions about inventory management, production planning, and marketing strategies for the next 6 months.

Month	Sales
Jan-23	139
Feb-23	152
Mar-23	165
Apr-23	178
May-23	191
Jun-23	204

Regression modeling: is used to analyze the relationship between two or more variables so as to identify the relationship between variables and predict future outcomes. E.g. in a housing dataset, it can be used to predict the price of a house based on its features such as location, number of bedrooms, and square footage.

Ex: Suppose we want to predict the price of a house based on its features such as the number of bedrooms, bathrooms, square footage, and age. We have a dataset of 1,000 houses with these features and their corresponding sale prices.

House	Bedrooms	Bathrooms	Sq. Footage	Age (years)	Sale Price ($1000s)
1	3	2	1500	10	300
2	4	3	2500	5	500
3	2	1	1000	20	200
…	…	…	…	…	…
1000	3	2	2000	15	400

We can use multiple linear regression, to predict the sale price of a house based on its features. We can formulate the regression model as follows:

Sale Price = β0 + β1 * Bedrooms + β2 * Bathrooms + β3 * Square Footage + β4 * Age + ε

where β0 is the intercept, β1 to β4 are the coefficients or parameters of the model, ε is the error term, and the other variables are the features of the house.

We can estimate the coefficients of the model using the least squares method, which minimizes the sum of squared errors between the predicted values and the actual values.

After we have estimated the coefficients, we can use the regression model to predict the sale price of a house with certain features.

Cluster analysis: is used to group data points into clusters based on their similarities so as to identify patterns and relationships within the data. E.g. in a marketing dataset, it can be used to group customers based on their purchasing behavior and identify target groups for marketing campaigns.

Ex: Suppose we have a dataset of 10 data points with two variables, x and y:

x	y
2	3
3	3
3	4
5	4
6	4
6	5
7	5
8	5
8	6
9	6

We want to group these data points into clusters based on their similarity in terms of x and y values. We can use a hierarchical clustering method with Euclidean distance as the similarity measure. Compute the pairwise distances between all data points:

	1	2	3	4	5	6	7	8	9
1	0.0	1.0	1.4	3.6	4.6	5.8	6.7	7.8	8.6
2	1.0	0.0	1.0	3.2	4.2	5.4	6.3	7.4	8.2
3	1.4	1.0	0.0	2.2	3.2	4.5	5.4	6.5	7.2
4	3.6	3.2	2.2	0.0	1.0	2.2	3.2	4.2	5.0
5	4.6	4.2	3.2	1.0	0.0	1.4	2.2	3.2	4.0
6	5.8	5.4	4.5	2.2	1.4	0.0	1.0	2.0	2.8
7	6.7	6.3	5.4	3.2	2.2	1.0	0.0	1.0	1.8
8	7.8	7.4	6.5	4.2	3.2	2.0	1.0	0.0	1.0
9	8.6	8.2	7.2	5.0	4.0	2.8	1.8	1.0	0.0

Then, we can use a hierarchical clustering algorithm to build a dendrogram that shows the clustering structure. We can use a single linkage method, which defines the distance between two clusters as the minimum distance between any two data points in the two clusters. The dendrogram would look like this:

```
  5     6
 / \   / \
1   2 3   4
     |
     7
    / \
   8   9
```

This dendrogram shows that the two most similar data points are 5 and 6, which are grouped into a cluster at the highest level. Then, the next most similar pair are 1 and 2, which are grouped into a cluster at the same level as 5 and 6.

Network analysis: is used to analyze the relationships between different nodes in a network so as to identify key nodes and relationships within the network. E.g. in a social network dataset, it can be used to identify key influencers and their relationships with other users.

Ex: Suppose we have a dataset of 6 people and their connections on a social network:

Person 1: 2, 3

Person 2: 1, 3, 4

Person 3: 1, 2, 4

Person 4: 2, 3, 5, 6

Person 5: 4, 6

Person 6: 4, 5

We can represent this dataset as a network, where each person is a node and each connection is an edge. We can create a graph visualization of this network using a tool like Gephi or Python's NetworkX library.

In this network, we can see the Person who has the most connections, and is therefore the most central node in the network.

We can use network analysis techniques to measure various properties of this network, such as centrality, clustering coefficient, and community structure. For example, we could calculate the degree centrality of each node to determine how many connections each person has. We could also calculate the clustering coefficient of each node to measure the density of connections within each person's immediate network. Additionally, we could use community detection algorithms to identify groups of people who are more closely connected to each other than to the rest of the network.

Step 8: Model Evaluation

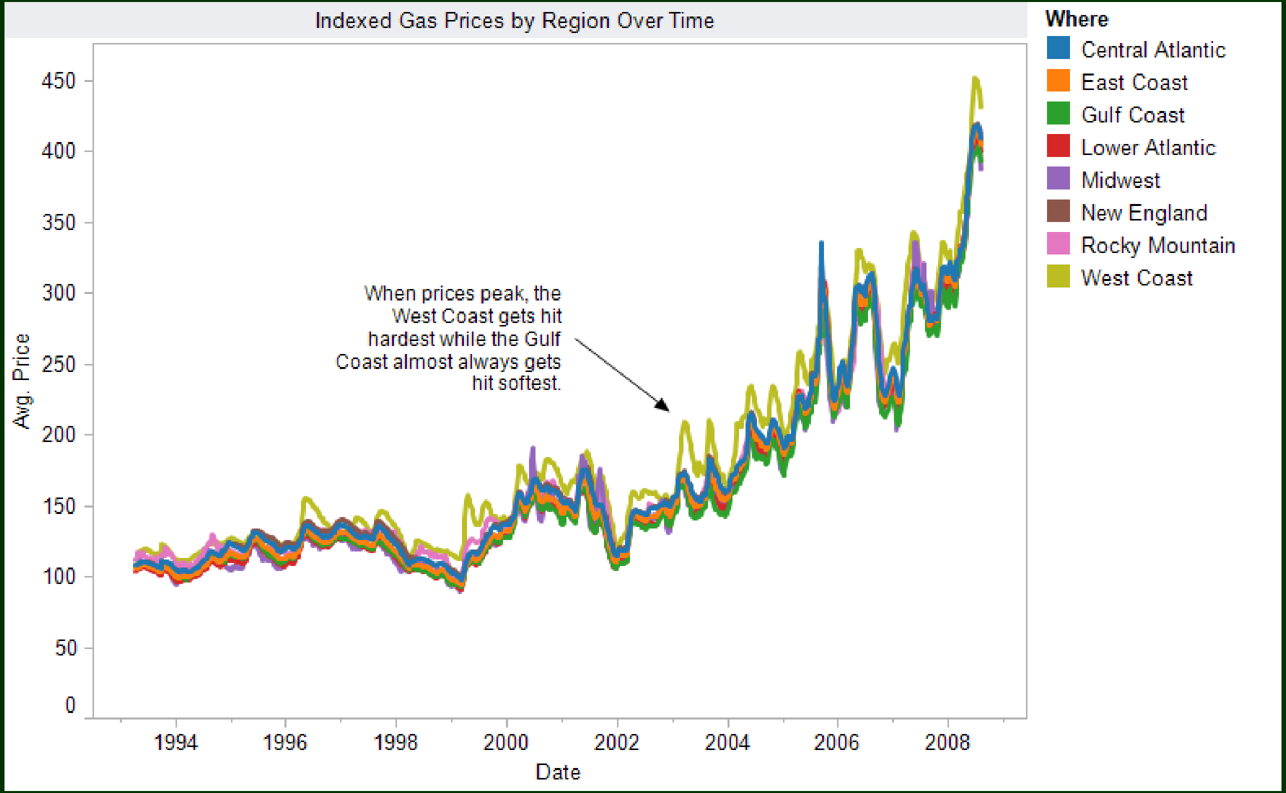
It can help assess the performance and accuracy of different models used for data analysis. Some examples of model evaluation are:

The data analyst can use metrics like MSE, R-squared, or RMSE to evaluate the models' performance and cross-validation techniques to estimate the models' performance on new, unseen data.

Visualization techniques like scatter plots or residual plots can be used to visually inspect the model's performance and identify any outliers or trends that may affect the model's performance. This information helps select the best model for the given problem and make informed decisions based on the insights gained from EDA.

Step 9: Communication

It involves extracting insights from data and communicating them to various stakeholders, including data analysts, business analysts, and decision-makers. The data analyst must communicate the results of their analysis in a clear and concise manner, using data visualizations such as graphs and charts to illustrate key findings. Collaboration with other stakeholders, such as business analysts or marketing teams, also requires effective communication to understand their needs and requirements, and tailor analysis and communication accordingly to ensure that the insights gained from EDA are relevant and actionable.


Indexed Gas Prices by Region Over Time
Where
🟦 Central Atlantic
🟧 East Coast
🟩 Gulf Coast
🟥 Lower Atlantic
🟪 Midwest
🟫 New England
Rocky Mountain
West Coast
When prices peak, the West Coast gets hit hardest while the Gulf Coast almost always gets hit softest.
Avg. Price
450
400
350
300
250
200
150
100
50
0
1994
1996
1998
2000
2002
2004
2006
2008
Date

Choosing the best Analysis Techniques

Factors to be considered:

Data Type: categorical, continuous, or a combination of both. E.g. descriptive statistics, such as mean and standard deviation, are appropriate for continuous data, while frequency distributions and contingency tables are appropriate for categorical data.

Research Question: e.g. if the research question is about identifying trends or patterns in the data, data visualization techniques such as histograms, scatter plots, and box plots may be useful. If the research question is about identifying relationships between variables, correlation analysis or regression analysis may be appropriate.

Sample Size: e.g. if the sample size is small, non-parametric tests may be more appropriate than parametric tests, which assume normality in the data distribution.

Data Quality: The quality of the data, including missing data, outliers, and errors, e.g. if the data contains outliers, robust statistical methods may be used to analyze the data.

Types of Multivariate Independence Techniques

- Principal Component Analysis (PCA)
- Cluster Analysis
- Factor Analysis
- Canonical Correlation Analysis
- Discriminant Analysis
- Multidimensional Scaling

Types of Multivariate Dependence Techniques

- Correlation Analysis
- Covariance Analysis
- Canonical Correlation Analysis
- Regression Analysis
- Factor Analysis
- Structural Equation Modeling

Canonical Correlation Analysis: Example

Suppose a firm surveyed a random sample of n = 50 of its employees to determine which factors influence sales performance. Two collections of variables were measured:

1. Sales Performance:
   - Sales Growth
   - Sales Profitability
   - New Account Sales
2. Test Scores as a Measure of Intelligence:
   - Creativity
   - Mechanical Reasoning
   - Abstract Reasoning
   - Mathematics

Factor Analysis: example

Factor analysis is used to identify underlying dimensions, or factors, that explain the variation in a set of observed variables.

Ex: Suppose we have a dataset of 10 variables that measure different aspects of a person's well-being, such as physical health, mental health, social support, and financial stability. We want to identify the underlying factors that are driving the variation in these variables.

We begin by performing a factor analysis on the data using a principal component analysis (PCA) method. The PCA method identifies the factors that explain the most variation in the data.

The factor analysis may reveal that three main factors explain the variation in the data, which we can label as follows:

Factor 1: Physical and Mental Health

- Physical health
- Mental health
- Sleep quality
- Diet quality

Factor 2: Social Support and Emotional Well-being

- Social support
- Emotional support
- Life satisfaction

Factor 3: Financial Stability and Security

- Income stability
- Financial security

Each factor is a linear combination of the original variables, with factor loadings indicating the strength of the relationship between each variable and the factor. For example, the factor loadings for Factor 1 may show that physical health and mental health have the strongest relationship with this factor, while sleep quality and diet quality have weaker relationships.

We can then use these factors as predictors in a regression model to understand how they are related to different outcomes, such as job performance or happiness. By using the factors instead of the original variables, we can simplify the model and reduce the risk of overfitting.

Factor analysis is useful because it can help to identify the underlying dimensions that are driving the variation in a dataset. This can make it easier to interpret results and identify meaningful patterns in the data. However, it is important to use appropriate methods for factor extraction and rotation and to ensure that the factors are interpretable and meaningful in the context of the research question.

Structural Equation Modeling: example

It is used to analyze complex relationships between multiple variables. It allows researchers to examine both the direct and indirect effects of variables on one another, and to test complex theoretical models that involve multiple variables.

Ex: Suppose we are interested in understanding the relationship between a person's socioeconomic status, their health behaviors (such as diet and exercise), and their risk of developing chronic diseases such as diabetes and heart disease. We collect data on 500 individuals and measure their income, education level, diet and exercise habits, and health outcomes.

We can then use SEM to test a theoretical model that describes the relationships between these variables.

The model might look like this:

Income and education level are exogenous variables that influence health behaviors

Health behaviors are endogenous variables that directly influence health outcomes

Income and education level also have indirect effects on health outcomes through their influence on health behaviors

We can use SEM to estimate the strength of these relationships and test the overall fit of the model to the data. If the model fits the data well, we can use it to make predictions about the relationships between these variables in a larger population.

SEM allows researchers to test complex theoretical models with multiple variables and to examine both direct and indirect effects of variables on one another. It also allows for the inclusion of measurement error and other sources of variability in the analysis.

Note: SEM requires a large sample size and a well-specified theoretical model. It can also be computationally intensive and requires specialized software and training to implement correctly. Nonetheless, SEM is a powerful tool for examining complex relationships between multiple variables and is widely used in social science, psychology, and other fields.

Ex 2: Suppose we want to test a theoretical model that suggests that academic achievement is influenced by multiple factors, including motivation, self-efficacy, and parental involvement.

Define latent variables (motivation, self-efficacy, parental involvement) and

Manifest variables (observed variables that can be measured) which will be used to operationalize each latent variable. For example, the manifest variables for motivation could be things like interest in school, engagement in class, and effort on assignments. Collect data from a sample of students and measure each of the manifest variables.

- Motivation <--- Interest in School
- Motivation <--- Engagement in Class
- Motivation <--- Effort on Assignments
- Self-efficacy <--- Confidence in Academic Abilities
- Self-efficacy <--- Perceived Control over Academic Success
- Parental Involvement <--- Communication with Parents about School
- Parental Involvement <--- Parental Monitoring of Schoolwork
- Academic Achievement <--- Motivation
- Academic Achievement <--- Self-efficacy
- Academic Achievement <--- Parental Involvement

The model suggests that motivation, self-efficacy, and parental involvement all have a direct effect on academic achievement, and that the manifest variables operationalizing each latent variable influence that latent variable.

Multi-variate ANOVA - example

ANOVA is a statistical technique for hypothesis testing to compare the differences between group means.

Ex: Suppose a pharmaceutical company wants to compare the efficacy of three different drugs (A, B, and C) in treating a certain medical condition. They collect data on the improvement in symptoms (measured on a scale from 0-10) of 100 patients who were randomly assigned to one of the three treatment groups. In addition, they also record the age, gender, and severity of the condition (measured on a scale from 0-100) for each patient.

Here's a summary of the data:

Treatment group A: improvement in symptoms (mean = 7.2), age (mean = 45 years), gender (60% female), severity (mean = 70)

Treatment group B: improvement in symptoms (mean = 6.5), age (mean = 47 years), gender (55% female), severity (mean = 75)

Treatment group C: improvement in symptoms (mean = 8.1), age (mean = 43 years), gender (45% female), severity (mean = 80)

To perform the MANOVA, the pharmaceutical company would first need to check if the assumptions of the analysis are met, such as normality, homogeneity of variances, and linearity of relationships between the dependent and independent variables. If the assumptions are met, they can then perform the MANOVA to determine if there are any significant differences in the improvement in symptoms across the three treatment groups while controlling for the effects of age, gender, and severity of the condition.

The output of the MANOVA analysis would include the multivariate F-test, which tests the null hypothesis that there are no significant differences in the improvement in symptoms across the three treatment groups while controlling for the effects of age, gender, and severity of the condition. If the multivariate F-test is significant, indicating that there are significant differences among the groups, then the pharmaceutical company can perform post-hoc tests, such as Bonferroni or Tukey's HSD, to determine which groups are significantly different from each other.

Multiple Discriminant Analysis

It is a statistical technique to identify the variables that best discriminate between two or more groups. It works by analyzing the relationships between the **predictor variables** and the **response variable** and seeks to identify a **linear combination of predictors** that maximizes the separation between the groups. It can be used for predicting group membership, identifying important predictor variables, and assessing the overall discriminative power of a set of predictors. It is a useful tool when working with multiple groups or categories of data.

- It is a regression technique to determine which particular group a data element belongs to.
- Multiple linear regression analysis is an extension of simple linear regression analysis.
- It is used for classifying data into groups.
- For more than 2 groups, Multiple Discriminant Analysis is used.

Let there be dataset with two groups A and B and two predictor variables, X1 and X2.

We want to identify the combination of X1 and X2 that best discriminates between the two groups. **T**he data is:

Group A:X1=1,X2=2; Group A:X1=2,X2=3; Group A:X1=3, X2=4

Group B:X1=5,X2=6; Group B:X1=6,X2=7; Group B:X1=7, X2=8

Find the mean vector and covariance matrix for each group.

Mean vectors are: Group A mean vector = [2, 3]

Group B mean vector = [6, 7]

Covariance matrices are:

Group A covariance matrix = [[1, 1], [1, 1]]

Group B covariance matrix = [[1, 1], [1, 1]]

Pooled within-group covariance matrix =
(3/4)*Group A covariancematrix + (3/4)*Group B covariancematrix
= [[1.5, 1.5], [1.5, 1.5]]

Coefficients for the discriminant function =
inv(Pooled within-group covariance matrix) *
(Group B mean vector - Group A mean vector) = [-4, -4]

The discriminant function is given by: Y = -4X1 - 4X2

To classify new observations, plug in their values for X1 and X2 into the discriminant function and compare the result to a threshold value.

For example, if Y < -20, the observation would be classified as Group A, and if Y > -20, it would be classified as Group B.

Conjoint Analysis

- It is a statistical technique used in market research to understand how people value different attributes (features) of a product or service. It is used in product design, pricing, and marketing research to help companies make better decisions about product development and marketing strategy.
- Conjoint Analysis works by presenting participants with a series of hypothetical product profiles or scenarios that vary in terms of different attributes and levels. For example, a hypothetical product profile might include information about the brand, price, size, color, and packaging of a product. Participants are asked to evaluate each product profile and rate their preference or likelihood to purchase the product.
- The results of these evaluations are used to estimate the relative importance of each attribute and level and to determine the ideal combination of attributes that will maximize customer satisfaction or purchase likelihood. This information can be used to guide product development decisions, pricing strategies, and marketing campaigns.
- It can be conducted using a variety of methods, including **full-profile conjoint analysis, adaptive conjoint analysis**, and **choice-based conjoint analysis.**

Ex: A company wants to understand how customers value different attributes of a burger, including patty size and toppings.

Conjoint Analysis involves presenting participants with two hypothetical burger profiles, each of which varies in terms of patty size and toppings. Participants are asked to choose which burger they would prefer to buy.

Burger 1: 1/4 lb. patty with lettuce, tomato, and ketchup

Burger 2: 1/3 lb. patty with bacon, cheese, and barbecue sauce

From the data, the company can estimate the relative importance of each attribute and level and determine the ideal combination of attributes that will maximize customer satisfaction or purchase likelihood.

Thanks

?